

QUADRATURE-BASED VECTOR FITTING: IMPLICATIONS FOR \mathcal{H}_2 SYSTEM APPROXIMATION

Z. DRMAČ*, S. GUGERCIN†, AND C. BEATTIE†

Abstract. Vector Fitting is a popular method of constructing rational approximants designed to fit given frequency response measurements. The original method, which we refer to as VF, is based on a least-squares fit to the measurements by a rational function, using an iterative reallocation of the poles of the approximant. We show that one can improve the performance of VF significantly, by using a particular choice of frequency sampling points and properly weighting their contribution based on quadrature rules that connect the least squares objective with an \mathcal{H}_2 error measure. Our modified approach, designated here as QuadVF, helps recover the original transfer function with better global fidelity (as measured with respect to the \mathcal{H}_2 norm), than the localized least squares approximation implicit in VF. We extend the new framework also to incorporate derivative information, leading to rational approximants that minimize system error with respect to a discrete Sobolev norm. We consider the convergence behavior of both VF and QuadVF as well, and evaluate potential numerical ill-conditioning of the underlying least-squares problems. We investigate briefly VF in the case of noisy measurements and propose a new formulation for the resulting approximation problem. Several numerical examples are provided to support the theoretical discussion.

Key words. least squares, frequency response, model order reduction, vector fitting, transfer function

AMS subject classifications. 34C20, 41A05, 49K15, 49M05, 93A15, 93C05, 93C15

1. Introduction. In many engineering applications, the dynamics that govern phenomena of interest may be inaccessible to direct modeling, yet there may be an abundance of accurate frequency response measurements available. In such cases, one may build up an empirical dynamical system model that fits the measured frequency response data. This empirical system may then be used as a surrogate to predict behavior or derive control strategies.

In other settings, one may have complete access to the underlying dynamical system of interest at least in principle (e.g., it may be an analytically derived computational model), however the full system may be a complex aggregate of many large subsystems, each perhaps representing diverse physics, and so it may be of such complexity that direct manipulation of the dynamical system is infeasible; potentially only simulation results would be available. Here, one may wish to capture the dominant dynamic features of the full aggregate system and realize them with a derived dynamical system (presumably of lower order) that can replicate the response characteristics of the full aggregate system. As before, this derived dynamical system may then be used as an efficient surrogate for the full system in contexts where performance is sensitive to model order.

A natural formulation of this task leads one to a data fitting problem using rational functions and this ultimately is our principal focus. For convenience, we assume that the system of interest is a single-input/single-output (SISO) linear time-invariant system associated with a transfer function, $H(s)$, that is unknown but accessible to sampling in the sense that measurements (magnitude and phase) of $H(s)$ at predetermined points, $s = \xi_1, \dots, \xi_\ell$ are available. Indeed, the values of $H(\xi_j)$, for $j = 1, \dots, \ell$ will be the only information presumed available for the system of interest. These values may have been obtained from experimentally measured amplitude and phase responses at $\xi_j = i 2\pi f_j$ associated with (real) driving frequencies, f_1, \dots, f_ℓ or they may have been extracted via simulation from a computational model.

We derive a dynamical system (or equivalently, its transfer function) by least squares (LS) data fitting: Denote by \mathcal{R}_r the set of proper rational functions of order r (i.e., with denominator having polynomial order r and numerator having polynomial order less than r). Fix ℓ sample points,

*Faculty of Science, Department of Mathematics, University of Zagreb, Bijenička 30, 10000 Zagreb, Croatia.

†Department of Mathematics, Virginia Polytechnic Institute and State University, 460 McBryde, Virginia Tech, Blacksburg, VA 24061-0123.

$\{\xi_j\}_1^\ell \in \mathbb{C}$, and weights, $\rho_j > 0$, for $j = 1, \dots, \ell$. The problem we address is stated succinctly as:

$$\begin{aligned} &\text{Find } H_r(s) \in \mathcal{R}_r \text{ such that } \sum_{j=1}^\ell \rho_j |H_r(\xi_j) - H(\xi_j)|^2 \longrightarrow \min \\ &(\text{i.e., for all } H_r(s) \in \mathcal{R}_r, \quad \sum_{j=1}^\ell \rho_j |H_r(\xi_j) - H(\xi_j)|^2 \leq \sum_{j=1}^\ell \rho_j |H_r(\xi_j) - H(\xi_j)|^2) \end{aligned} \quad (1.1)$$

Typically, all $\rho_j = 1$ (the “unweighted” case). We will be interested in strategies that take advantage of other choices for ρ_j (which may lead to particular choices for ξ_j , as well). Rational data fitting strategies brought into the service of systems identification in this way have a long history going back at least to Kalman [35], who computed a best least squares fit with sampled input and output data using rational functions of the form $\sum_{j=1}^r a_j z^{-j} / (1 + \sum_{j=1}^r b_j z^{-j})$ (in the z -transform domain).

Levy [40] considered (1.1), taking the rational approximants, H_r , to be in polynomial form:

$$H_r(s) = \frac{n(s)}{d(s)} \text{ with } n(s) = \sum_{j=0}^{r-1} \alpha_j s^j \text{ and } d(s) = 1 + \sum_{j=1}^r \beta_j s^j. \quad (1.2)$$

Since the set of rational functions, \mathcal{R}_r , is not an affine set (indeed, not even convex), (1.1) is both nonlinear and nonconvex, leading possibly to a host of local minima. Noting first that

$$\sum_{j=1}^\ell |H_r(\xi_j) - H(\xi_j)|^2 = \sum_{i=1}^\ell \frac{1}{|d(\xi_i)|^2} |n(\xi_i) - d(\xi_i)H(\xi_i)|^2, \quad (1.3)$$

Levy proposed replacing (1.1) with the simpler problem of minimizing $\sum_{i=1}^\ell |n(\xi_i) - d(\xi_i)H(\xi_i)|^2$; an LS problem which is linear in the coefficients $\{\alpha_j\}$, $\{\beta_j\}$. Sanathanan and Koerner [49] argued against this tactic and provided a convincing example that such a simplification is problematic. They suggested an iterative approach for solving (1.1) that used Levy’s simplification as a first step.

We refer to this approach as *SK iteration* and describe two equivalent formulations of it in §2. One of these formulations leads to a particularly interesting refinement, introduced by Gustavsen and Semlyen [30] under the name *Vector Fitting* (VF). We describe VF in §2 and make some observations that will contribute to our analysis of it in §3. Overall, VF has been a great success, with more than 700 citations and a wide spectrum of applications. Many authors have applied, modified, and analyzed VF, see e.g. [29], [32], [18], [17], [20], [19]. Our motivation for studying VF came initially from a desire to articulate the relationship between VF and optimal rational approximation, in particular, with \mathcal{H}_2 -optimal model order reduction. We set the stage for this in §3 where we review some basic results related to \mathcal{H}_2 -optimal rational approximation. We show that a small VF fitting error does not necessarily correspond to small approximation error in the \mathcal{H}_2 or \mathcal{H}_∞ norm. This observation motivates the developments of §3.2, §3.3, where we show that particular choices of sampling points and weights, as dictated by suitable quadrature formulae, may significantly improve the performance of VF. The key innovation here lies in reformulating the problem essentially as an approximation problem in a normed function space instead of as an algebraic LS problem.

Some implementation details are provided in §4. Formal mathematical justification of mirroring unstable nodes in VF is given in §4.1. In §4.2, we use numerical examples to illustrate the complexity of the theoretically open problem of the convergence of VF iterations. In §4.3, we discuss the important issue of high condition numbers of the matrices used in VF, and introduce a regularized LS version of VF. The behavior of VF in the case of noisy data is analyzed in §5, where we show that VF will asymptotically and implicitly solve a structured total least squares problem in computing the coefficients. This goes some distance in explaining the robustness observed in VF.

In recent years, the Loewner framework, initially introduced by Mayo and Antoulas [41] and further extended in [4, 38, 44], has emerged as a powerful, effective and numerically efficient method to construct rational approximants directly from frequency domain measurements. Our major focus

in this paper is the rational least-squares approximation produced by VF; to investigate VF from an optimal approximation perspective, to offer improvement based on this analysis and to examine several computational issues. A comparison of VF with the Loewner framework and related approaches is natural to consider however it will not be considered here.

2. The Sanathanan-Koerner Iteration and Vector Fitting.

2.1. SK iteration. Sanathanan and Koerner [49] noted that minimizing the objective function $\sum_{i=1}^{\ell} |n(\xi_i) - d(\xi_i)H(\xi_i)|^2$ instead of (1.3) could produce quite different outcomes since $|d(\xi_i)|$ could vary over a wide range of magnitudes. They offered an alternative approach through the iterative adjustment of the LS weights:

$$\text{Starting with } d^{(0)}(s) \equiv 1, \text{ solve successively for } k = 0, 1, 2, \dots$$

$$\sum_{i=1}^{\ell} \left| \frac{n^{(k+1)}(\xi_i) - d^{(k+1)}(\xi_i)H(\xi_i)}{d^{(k)}(\xi_i)} \right|^2 \rightarrow \min. \quad (2.1)$$

We will refer to this process as “SK iteration.”

Polynomial Representation. Using the polynomial representation of $H_r(s)$ in (1.2), one may reformulate (2.1) as a weighted LS problem (following [49]):

$$\|\Delta^{(k)}(\mathcal{B}y^{(k+1)} - h)\|_2 \rightarrow \min \quad (2.2)$$

where the optimization parameters are $y^{(k+1)} = (\alpha_0^{(k+1)} \alpha_1^{(k+1)} \dots \alpha_{r-1}^{(k+1)} \beta_1^{(k+1)} \beta_2^{(k+1)} \dots \beta_r^{(k+1)})^T$, while

$$\mathcal{B} = \begin{pmatrix} 1 & \xi_1 & \xi_1^2 & \dots & \xi_1^{r-1} & -H(\xi_1)\xi_1 & -H(\xi_1)\xi_1^2 & \dots & -H(\xi_1)\xi_1^r \\ 1 & \xi_2 & \xi_2^2 & \dots & \xi_2^{r-1} & -H(\xi_2)\xi_2 & -H(\xi_2)\xi_2^2 & \dots & -H(\xi_2)\xi_2^r \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & \xi_{\ell-1} & \xi_{\ell-1}^2 & \dots & \xi_{\ell-1}^{r-1} & -H(\xi_{\ell-1})\xi_{\ell-1} & -H(\xi_{\ell-1})\xi_{\ell-1}^2 & \dots & -H(\xi_{\ell-1})\xi_{\ell-1}^r \\ 1 & \xi_{\ell} & \xi_{\ell}^2 & \dots & \xi_{\ell}^{r-1} & -H(\xi_{\ell})\xi_{\ell} & -H(\xi_{\ell})\xi_{\ell}^2 & \dots & -H(\xi_{\ell})\xi_{\ell}^r \end{pmatrix}, \quad h = \begin{pmatrix} H(\xi_1) \\ H(\xi_2) \\ \vdots \\ H(\xi_{\ell-1}) \\ H(\xi_{\ell}) \end{pmatrix}, \quad (2.3)$$

and $\Delta^{(k)} = \text{diag} \left(\frac{1}{|d^{(k)}(\xi_j)|} \right)_{j=1}^{\ell}.$

The sequence of LS solutions, $y^{(k)}$, yields polynomial coefficients for the sequence of numerators, $n^{(k)}(s)$, and denominators, $d^{(k)}(s)$, of $H_r^{(k)}(s)$ (as in (1.2)). If the denominator sequence, $d^{(k)}(s)$, converges, then so does the numerator sequence, $n^{(k)}(s)$, and so the SK iteration (2.1) produces a system $H_r(s)$ that may be expected to be a locally optimal solution to (1.1).

Barycentric representation. The rational function $H_r(s)$ in (1.2) can be represented alternatively in barycentric form, which happens here to be both elegant and useful. We develop this by expressing the numerator and the denominator in a Lagrange interpolating basis: Pick an arbitrary set of mutually distinct scalars $\lambda_0, \lambda_1, \dots, \lambda_r$ (“interpolation points”) and define the nodal polynomial $\omega_r(s) = \prod_{k=1}^r (s - \lambda_k)$ (notice λ_0 is excluded). Then,

$$n(s) = \omega_r(s) \sum_{j=1}^r \frac{w_j n(\lambda_j)}{s - \lambda_j} \quad \text{and} \quad d(s) = \omega_r(s) \left(\alpha + \sum_{j=1}^r \frac{w_j d(\lambda_j)}{s - \lambda_j} \right),$$

where $w_j = 1 / \prod_{k \neq j} (\lambda_j - \lambda_k)$ enforces interpolation of $n(s)$, and hence $H_r(s)$, at $s = \lambda_j$ for $j = 1, \dots, r$ and choosing $\alpha = \frac{d(\lambda_0)}{\omega_r(\lambda_0)} - \sum_{j=1}^r \frac{d(\lambda_j)w_j}{\lambda_0 - \lambda_j}$ then enforces interpolation of H_r also at $s = \lambda_0$. As long as $d(s)$ has polynomial degree r , then $\alpha \neq 0$. Define $\phi_j = \frac{w_j}{\alpha} n(\lambda_j)$ and $\varphi_j = \frac{w_j}{\alpha} d(\lambda_j)$, so

$$H_r(s) = \frac{\sum_{j=1}^r \frac{\phi_j}{s - \lambda_j}}{1 + \sum_{j=1}^r \frac{\varphi_j}{s - \lambda_j}} = \frac{\tilde{n}(s)}{\tilde{d}(s)} \quad \text{with} \quad \begin{cases} \tilde{n}(s) = \sum_{j=1}^r \frac{\phi_j}{s - \lambda_j}, \text{ and} \\ \tilde{d}(s) = 1 + \sum_{j=1}^r \frac{\varphi_j}{s - \lambda_j}. \end{cases} \quad (2.4)$$

We may now use ϕ_j, φ_j as optimization parameters in each step of the SK iteration (2.1). Indeed, for a given set of interpolation points, $\lambda_1, \dots, \lambda_r$, observation points ξ_1, \dots, ξ_ℓ , and system observations $H(\xi_1), \dots, H(\xi_\ell)$, the parameters $\phi_j^{(k)}, \varphi_j^{(k)}$ describe $H_r^{(k)}(s) = \frac{\tilde{n}^{(k)}(s)}{\tilde{d}^{(k)}(s)}$ in the k th step of (2.1), replacing $n^{(k)}$ and $d^{(k)}$ in (2.1) with

$$\tilde{n}^{(k)}(s) = \sum_{j=1}^r \frac{\phi_j^{(k)}}{s - \lambda_j} \quad \text{and} \quad \tilde{d}^{(k)}(s) = 1 + \sum_{j=1}^r \frac{\varphi_j^{(k)}}{s - \lambda_j}, \quad (2.5)$$

respectively. Now, $\phi_j^{(k)}, \varphi_j^{(k)}$ are determined by solution of the successive least squares problems

$$\|\Delta^{(k)}(\mathcal{A}x^{(k+1)} - h)\|_2 \rightarrow \min, \quad k = 0, 1, 2, \dots, \quad (2.6)$$

where the unknowns now are $x^{(k+1)} = (\phi_1^{(k+1)} \phi_2^{(k+1)} \dots \phi_r^{(k+1)} \varphi_1^{(k+1)} \varphi_2^{(k+1)} \dots \varphi_r^{(k+1)})^T$ and

$$\mathcal{A} = \begin{pmatrix} \frac{1}{\xi_1 - \lambda_1} & \frac{1}{\xi_1 - \lambda_2} & \dots & \frac{1}{\xi_1 - \lambda_r} & \frac{-H(\xi_1)}{\xi_1 - \lambda_1} & \frac{-H(\xi_1)}{\xi_1 - \lambda_2} & \dots & \frac{-H(\xi_1)}{\xi_1 - \lambda_r} \\ \frac{1}{\xi_2 - \lambda_1} & \frac{1}{\xi_2 - \lambda_2} & \dots & \frac{1}{\xi_2 - \lambda_r} & \frac{-H(\xi_2)}{\xi_2 - \lambda_1} & \frac{-H(\xi_2)}{\xi_2 - \lambda_2} & \dots & \frac{-H(\xi_2)}{\xi_2 - \lambda_r} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{1}{\xi_{\ell-1} - \lambda_1} & \frac{1}{\xi_{\ell-1} - \lambda_2} & \dots & \frac{1}{\xi_{\ell-1} - \lambda_r} & \frac{-H(\xi_{\ell-1})}{\xi_{\ell-1} - \lambda_1} & \frac{-H(\xi_{\ell-1})}{\xi_{\ell-1} - \lambda_2} & \dots & \frac{-H(\xi_{\ell-1})}{\xi_{\ell-1} - \lambda_r} \\ \frac{1}{\xi_\ell - \lambda_1} & \frac{1}{\xi_\ell - \lambda_2} & \dots & \frac{1}{\xi_\ell - \lambda_r} & \frac{-H(\xi_\ell)}{\xi_\ell - \lambda_1} & \frac{-H(\xi_\ell)}{\xi_\ell - \lambda_2} & \dots & \frac{-H(\xi_\ell)}{\xi_\ell - \lambda_r} \end{pmatrix}. \quad (2.7)$$

Note h and $\Delta^{(k)}$ are as defined in (2.3), with $\tilde{d}^{(k)}(s)$ as given in (2.5) replacing $d^{(k)}(s)$ in $\Delta^{(k)}$.

Equivalence of the representations. It is straightforward to see that both (2.6)-(2.7) and (2.2)-(2.3) are simply different representations of the same iteration step described in (2.1), the key difference being that $H_r(s)$ is expressed with respect to different bases. Note that \mathcal{B} in (2.2)-(2.3) depends solely on the complex frequency points, ξ_i , at which the system is observed, while \mathcal{A} in (2.6)-(2.7) depends both on those observation points, $\{\xi_i\}$ and on auxiliary interpolation points, $\{\lambda_j\}$. The interpolation points (λ s) used in the definition of \mathcal{A} have been chosen arbitrarily; they serve just to fix a particular barycentric representation, and remain constant throughout the iteration.

Interestingly, if the interpolation points used in the definition of \mathcal{A} are chosen to be the r th roots of unity, $\lambda_j = \bar{\omega}^{j-1}$ with $\omega = e^{i(2\pi/r)}$, then one can show that the SK iterations in (2.2) and (2.6) are related via the r -dimensional discrete Fourier Transform, $\mathbb{F} \in \mathbb{C}^{r \times r}$ with $\mathbb{F}_{ij} = \frac{\omega^{(i-1)(j-1)}}{\sqrt{r}}$. More precisely, solving $\|\Delta^{(k)}(\mathcal{B}y^{(k+1)} - h)\|_2 \rightarrow \min$ in the usual polynomial basis is equivalent to solving $\|\tilde{\Delta}^{(k)}(\mathcal{A}\tilde{x}^{(k+1)} - D_1^{-1}h)\|_2 \rightarrow \min$ with a particular choice of barycentric representation, and the two solutions are related by

$$y^{(k+1)} = \mathbb{F} D_2 \tilde{x}^{(k+1)}, \quad \text{where} \quad (D_1)_{ii} = \frac{\xi_i^n - 1}{\sqrt{r}}, \quad (D_2)_{jj} = \frac{1}{\omega^{j-1}}. \quad (2.8)$$

Each of the iterative processes described in (2.2)-(2.3) and in (2.6)-(2.7) are concrete realizations of (2.1), and as such, they each are driven by successive updates of the weighting factors $\Delta^{(k)}$. As the weighting factors, $\Delta^{(k)}$, change, so too do the denominators of the approximants $H_r^{(k)}(s) = \frac{n^{(k)}(s)}{d^{(k)}(s)}$ and, in particular, the poles of $H_r^{(k)}(s)$ will change from step to step. No constraint has been imposed that guarantees these poles remain in the left half-plane, and so it may happen that a minimizing solution to (2.1) produces an unstable system, an outcome that would generally be viewed as unsatisfactory. Thus, as a practical matter, it is necessary additionally to monitor the zeros of the denominators, $d^{(k)}(s)$, and, perhaps on occasion, intercede to repair unstable poles as they emerge (e.g., by reflecting them across the imaginary axis back into the left half-plane). *Vector Fitting*, as we see next, also uses this information to determine an advantageous representation for the next step in (2.1).

2.2. Vector Fitting (VF) [30]. Since the choice of the interpolation points in the SK iteration only determines a particular barycentric representation for rational functions, one is free to change the λ_j at every step. The original formulation of *Vector Fitting* as introduced by Gustavsen and Semlyen [30] takes advantage of this flexibility and cleverly updates the interpolation points in the course of the iteration. In addition to providing more accurate rational approximants and generally providing greater stability and better performance than the SK iteration, this dynamic updating of the interpolation points achieves other useful goals, as explained below and in §5.

Suppose now that the interpolation points depend on k and denote them by $\lambda_j^{(k)}$; we define $\mathcal{A}^{(k)} \equiv \mathcal{A}(\boldsymbol{\lambda}^{(k)})$ to be \mathcal{A} as defined in (2.7), but with λ_j replaced by $\lambda_j^{(k)}$. After the k -th step of the iteration, VF assigns $\lambda_j^{(k+1)}$ to be the zeros of $\tilde{d}^{(k)}(s)$ in (2.5):

$$\tilde{d}^{(k)}(s) = 1 + \sum_{j=1}^r \frac{\varphi_j^{(k)}}{s - \lambda_j^{(k)}} = \frac{\prod_{j=1}^r (s - \lambda_j^{(k+1)})}{\prod_{j=1}^r (s - \lambda_j^{(k)})}. \quad (2.9)$$

Then, the goal of (2.6) becomes the minimization of

$$\begin{aligned} \|\Delta^{(k)}(\mathcal{A}^{(k)}x^{(k+1)} - h)\|_2^2 &= \sum_{i=1}^{\ell} \frac{1}{|\tilde{d}^{(k)}(\xi_i)|^2} \left| \sum_{j=1}^r \frac{\phi_j^{(k+1)}}{\xi_i - \lambda_j^{(k)}} - H(\xi_i) \left(1 + \sum_{j=1}^r \frac{\varphi_j^{(k+1)}}{\xi_i - \lambda_j^{(k)}} \right) \right|^2 \\ &= \sum_{i=1}^{\ell} \left| \frac{\prod_{j=1}^r (\xi_i - \lambda_j^{(k)})}{\prod_{j=1}^r (\xi_i - \lambda_j^{(k+1)})} \right|^2 \left| \frac{\tilde{p}^{(k+1)}(\xi_i)}{\prod_{j=1}^r (\xi_i - \lambda_j^{(k)})} - H(\xi_i) \frac{\tilde{q}^{(k+1)}(\xi_i)}{\prod_{j=1}^r (\xi_i - \lambda_j^{(k)})} \right|^2 \end{aligned} \quad (2.10)$$

where $\tilde{p}^{(k+1)}$ and $\tilde{q}^{(k+1)}$ are, respectively, polynomials of degree $r-1$ and r . Continuing with similar algebraic manipulations, one obtains

$$\begin{aligned} \|\Delta^{(k)}(\mathcal{A}^{(k)}x^{(k+1)} - h)\|_2^2 &= \sum_{i=1}^{\ell} \left| \frac{\tilde{p}^{(k+1)}(\xi_i)}{\prod_{j=1}^r (\xi_i - \lambda_j^{(k+1)})} - H(\xi_i) \frac{\tilde{q}^{(k+1)}(\xi_i)}{\prod_{j=1}^r (\xi_i - \lambda_j^{(k+1)})} \right|^2 \\ &= \sum_{i=1}^{\ell} \left| \sum_{j=1}^r \frac{\tilde{\phi}_j^{(k+1)}}{\xi_i - \lambda_j^{(k+1)}} - H(\xi_i) \left(1 + \sum_{j=1}^r \frac{\tilde{\varphi}_j^{(k+1)}}{\xi_i - \lambda_j^{(k+1)}} \right) \right|^2 \\ &= \|\mathcal{A}^{(k+1)}\tilde{x}^{(k+1)} - h\|_2^2, \end{aligned} \quad (2.11)$$

where $\tilde{x}^{(k+1)} = \left(\tilde{\phi}_1^{(k+1)} \tilde{\phi}_2^{(k+1)} \dots \tilde{\phi}_r^{(k+1)} \tilde{\varphi}_1^{(k+1)} \tilde{\varphi}_2^{(k+1)} \dots \tilde{\varphi}_r^{(k+1)} \right)^T$ with $\tilde{\phi}_j^{(k+1)}$ and $\tilde{\varphi}_j^{(k+1)}$ as defined in (2.11). Thus, one step of VF corresponds to solving the least squares problem

$$\|\mathcal{A}^{(k+1)}\tilde{x}^{(k+1)} - h\|_2 \rightarrow \min, \quad k = 0, 1, 2, \dots \quad (2.12)$$

This is an unweighted LS step using an updated barycentric representation of $H_r(s)$ based on $\boldsymbol{\lambda}^{(k+1)}$ and with the coefficient matrix $\mathcal{A}^{(k+1)} = \mathcal{A}(\boldsymbol{\lambda}^{(k+1)})$; effectively, one step of the SK iteration with unity weighting. For this reason, VF may be thought of as a representation of SK iteration in a well-chosen basis [32]. One of the points we make in this paper is that VF is more than that.

The scaling that underlies the SK iteration is implicit in (2.12) and provides a critical correction to the approximation metric when close to the true minimizer. However, when the approximant, $H_r^{(k)}$ is far from the true minimizer, that same scaling may inflict severe damage on the early evolution of the iterations, leading subsequent iterates to an unsatisfactory final approximant (cf. [49]). This

makes the performance of the SK iteration (and hence also the VF iteration) potentially sensitive to the quality of initialization.

Since VF assigns $\lambda_j^{(k+1)}$ to be the zeros of $\tilde{d}^{(k)}(s)$, the poles of $\tilde{d}^{(k+1)}(s)$ will be zeros of $\tilde{d}^{(k)}(s)$ and in the limit, assuming convergence, pole-zero cancelation occurs. If the interpolation points, $\lambda_j^{(k)}$, converge to finite values as $k \rightarrow \infty$ then from (2.9), $\tilde{d}^{(k)}(s) \rightarrow 1$ and, in the limit, $\tilde{n}^{(k)}(s)$ will give the final rational approximant in the pole-residue representation. However, theoretical convergence of VF is still an open problem, and a careful justification of the stopping criterion (e.g. using backward error analysis) is also lacking. We address these issues in more detail in §4.2.

3. Vector Fitting and Discrete \mathcal{H}_2 Approximation.

3.1. \mathcal{H}_2 approximation. Let $\mathcal{H}_2(\mathbb{C}_+)$ denote the vector space of complex functions, $H(s)$, that are analytic in the open right-half plane, $\mathbb{C}_+ = \{s \equiv x + iy \in \mathbb{C} : x > 0\}$, such that $\sup_{x>0} \int_{-\infty}^{+\infty} |H(x + iy)|^2 dy < \infty$. $\mathcal{H}_2(\mathbb{C}_+)$ is a Hilbert space endowed with an inner product

$$\langle G, H \rangle_{\mathcal{H}_2} = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \overline{H(i\omega)} G(i\omega) d\omega, \text{ and norm } \|G\|_{\mathcal{H}_2} = \sqrt{\langle G, G \rangle_{\mathcal{H}_2}}. \quad (3.1)$$

The boundary operator isometry $\mathcal{T} : \mathcal{H}_2(\mathbb{C}_+) \rightarrow L_2(i\mathbb{R})$, $\mathcal{T}[H](i\omega) = \lim_{x \downarrow 0} H(x + i\omega)$, identifies H with its boundary function, $\mathcal{H}_2(\mathbb{C}_+) \cong \text{Range}(\mathcal{T}) \subset L_2(i\mathbb{R})$. If G and H are strictly proper rational functions representing transfer functions of real stable linear time invariant dynamical systems then $G, H \in \mathcal{H}_2(\mathbb{C}_+)$, and we have in addition,

$$\langle G, H \rangle_{\mathcal{H}_2} = \langle H, G \rangle_{\mathcal{H}_2} = \frac{1}{2\pi} \int_{-\infty}^{+\infty} H(-i\omega) G(i\omega) d\omega \quad \text{and} \quad G(s) = \frac{1}{2\pi} \int_{\mathbb{R}} \frac{G(i\omega)}{s - i\omega} d\omega.$$

If H_r is an \mathcal{H}_2 -optimal r th order rational approximation to a given $H(s) \in \mathcal{H}_2$, then it must be a Hermite interpolant of $H(s)$ in the following sense: Suppose

$$H_r(s) = \sum_{i=1}^r \frac{\phi_i}{s - \lambda_i} = \underset{\substack{\text{order } \tilde{H}_r \leq r \\ \tilde{H}_r \text{ stable}}}{\text{argmin}} \|H - \tilde{H}_r\|_{\mathcal{H}_2}.$$

Then,

$$H(-\lambda_j) = H_r(-\lambda_j) \quad \text{and} \quad H'(-\lambda_j) = H'_r(-\lambda_j), \quad \text{for } j = 1, 2, \dots, r; \quad (3.2)$$

$H_r(s)$ is a Hermite interpolant to $H(s)$ at the mirror images of its own poles reflected across the imaginary axis [27, 43]. These optimal interpolation points, $\{-\lambda_i\}_{i=1}^r$, evidently depend on the poles of the optimal approximant that is sought, so they are not known *a priori*. The *Iterative Rational Krylov Algorithm* (IRKA) of Gugercin et al. [27] is a numerically effective iterative correction process that systematically enforces these necessary conditions for optimality.

The original formulation of IRKA described in [27] requires access to a first-order state-space realization for $H(s)$: $H(s) = \mathbf{C}(s\mathbf{E} - \mathbf{F})^{-1}\mathbf{B}$. By employing a Loewner-matrix framework introduced by Mayo and Antoulas [41], Beattie and Gugercin [8] relaxed this requirement; one only needs the ability to evaluate $H(s)$ for $s \in \mathbb{C}$ in order to obtain (locally) \mathcal{H}_2 -optimal rational approximants to $H(s)$. This has allowed effective data-driven \mathcal{H}_2 -optimal system approximation for a much larger class of functions, including many that are not necessarily rational such as arise with delay systems. For more details on optimal \mathcal{H}_2 approximation, see [3, 27, 43, 51, 55] and references therein.

Notably, the data required to run the Loewner-IRKA approach of [8] is similar to what is required for VF but with one important difference, neither the number nor the location of the points of evaluation of $H(s)$ is known in advance for the Loewner-IRKA approach. This is in contrast to VF

where a predetermined number of $H(s)$ evaluations are computed (or provided by simulation) at the beginning and the rest of the process does not require any new $H(s)$ evaluations. This, of course, comes with the disadvantage that the resulting approximation due to VF will fit only the sampling of $H(s)$ that had been acquired and so it ultimately may be a poor approximation to $H(s)$ with respect to an \mathcal{H}_2 or \mathcal{H}_∞ measure.

3.2. Reformulating Vector Fitting as Discrete \mathcal{H}_2 minimization. VF is widely recognized as a very effective tool in creating rational approximants that fit frequency-sampled functions. How best to organize the necessary frequency sampling is not discussed in general and seems governed more by expedience with just a few general guidelines. For example, in the discussion portion of [30], the authors offer the heuristic "The samples should be chosen so densely that the frequency response is fully resolved." They go on to recommend having at least as many samples as poles (r) and, in turn, at least twice as many poles as there are peaks in the frequency response. These are useful guidelines, yet clearly they do not (nor are they intended to) cover all cases of interest: for example, high modal densities can obscure resonances. Moreover, if significant expense is associated with obtaining each frequency sample, then one is motivated to reduce sampling density and one may be forced to enter the gray area between a sampling density that "fully resolves" the frequency response and one that may leave it "unresolved." Indeed, certain application settings may not allow sufficient sampling density to resolve fully the frequency response and one wishes then to maximize the effectiveness of parsimonious sampling strategies.

EXAMPLE 3.1. Consider the FOM Model from the NICONET Benchmark collection [14]. The model $H(s)$ has order $n = 1006$, yet the frequency response has only three obvious peaks, between 8 Hz and 160 Hz. We create a rational approximant of order $r = 12$ using VF with $\ell = 40$ frequency sampling points f_i , logarithmically spaced between 10^{-3} and 10^3 . VF was very effective in producing a rational approximant with an excellent goodness-of-fit; the relative least-squares residual was 2.75×10^{-4} . However, this did not mean a high-fidelity model was obtained: indeed, the corresponding relative \mathcal{H}_2 error was only 1.78×10^{-1} and much better models of the same order can be obtained easily. Applying IRKA to the same system produced a model of the same order, but with a relative \mathcal{H}_2 error of only 1.92×10^{-4} , an approximation that is virtually indistinguishable from the original. Not surprisingly, this greater accuracy came at a somewhat greater cost: On this example, IRKA took 5 iterations to converge. Every iteration step required twelve $H(s)$ evaluations and twelve $H'(s)$ evaluations. However, the twelve interpolation points comprised 3 complex conjugate pairs and 6 real points in each iteration, so every iteration required only nine independent $H(s)$ and nine independent $H'(s)$ evaluations, netting a total of $\ell = 45$ $H(s)$ and $\ell = 45$ $H'(s)$ evaluations. The main point to note in this regard is not so much the number of function/derivative evaluations — it is often the case that function and derivative evaluations can be combined so the net computational effort, both in this case and in general, is typically far less than twice what is required just for function evaluations. Rather, one should note that with IRKA (and in contrast with VF), one cannot anticipate exactly *where* these function evaluations will occur beforehand.

Our goal is to bring the achievable accuracy of VF more in line with what IRKA can provide, without sacrificing its attractive computational features. We find that by interpreting the VF objective function of (1.1) as a discretization of an \mathcal{H}_2 error measure, remarkably effective sampling strategies may be developed systematically through numerical quadrature. The general approach that we will take in the sequel arrives at a vector fitting formulation (1.1) by approximating the \mathcal{H}_2 error with an appropriate quadrature rule. This will lead us to minor modifications of VF that we find often dramatically improves its quality of approximation.

3.3. Effective Sampling Points via Quadrature. Approximating the \mathcal{H}_2 error measure with a quadrature rule leads one to consider approximations of the form

$$\int_{-\infty}^{+\infty} |H(i\omega) - H_r(i\omega)|^2 d\omega \approx \sum_{j=1}^{\ell} \rho_j^2 |H(\xi_j) - H_r(\xi_j)|^2 + \rho_+^2 M_+[|H - H_r|^2] + \rho_-^2 M_-[|H - H_r|^2] \quad (3.3)$$

where $M_{\pm}[G]$ are linear functionals of G that capture information about behavior at $\pm\infty$. Note that if $\rho_+ = \rho_- = 0$, with all other $\rho_j = 1$, and if sampling nodes, ξ_j , are chosen to be equidistant and in complex conjugate pairs, then we recover the usual VF objective function which then can be understood as a composite trapezoid quadrature rule for the integral in (3.3), giving the \mathcal{H}_2 error.

Of course, the trapezoid rule will not be an optimal choice of quadrature rule in most cases and many, much more effective options are easily formulated, many of which involve first mapping the unbounded domain of integration, $(-\infty, \infty)$, to a finite interval, often either $(-1, 1)$ or $(0, \pi)$, and then applying a high accuracy quadrature rule. We focus on a quadrature rule developed by Boyd [13], which is related to Clenshaw-Curtis quadrature and chosen here for its simplicity. Many options of this sort may be considered; our main goal is to illustrate the potential of this approach without overburdening the reader with technicalities.

Adapted to our setting, the Boyd/Clenshaw-Curtis (B/CC) formula [13] is

$$\begin{aligned} \|H(s)\|_{\mathcal{H}_2}^2 &= \int_{-\infty}^{+\infty} |H(i\omega)|^2 d\omega = \int_0^{\pi} \frac{L}{\sin^2 t} |H(iL \cot t)|^2 dt \\ &\approx \sum_{j=1}^{\ell} \frac{L\pi}{(\ell+1)\sin^2 t_j} |H(iL \cot t_j)|^2 + \frac{\pi}{2L(\ell+1)} (|M_+[H]|^2 + |M_-[H]|^2). \end{aligned} \quad (3.4)$$

where $L > 0$ is a freely chosen scaling parameter, $t_j = \frac{j\pi}{\ell+1}$, for $j = 1, \dots, \ell$, and

$$\begin{aligned} M_+[G] &= \lim_{\omega \rightarrow \infty} i\omega G(i\omega) = \lim_{t \rightarrow 0^+} \frac{G(iL \cot t)}{\sin(t)} \cdot iL \\ M_-[G] &= \lim_{\omega \rightarrow -\infty} i\omega G(i\omega) = \lim_{t \rightarrow \pi^-} \frac{G(iL \cot t)}{\sin(t)} \cdot iL \end{aligned} \quad (3.5)$$

For example, if $H(s)$ is a strictly proper transfer function with realization, $H(s) = \mathbf{C}(s\mathbf{I} - \mathbf{F})^{-1}\mathbf{B}$, then $M_+[H] = M_-[H] = \mathbf{CB}$.

The choice of L can influence greatly the accuracy of this quadrature rule. Notice that as the value of L decreases, the quadrature nodes are drawn towards the origin with diminished weight, while contributions at $\pm\infty$ have increased weight to compensate. Boyd [12] observed that when integrands are entire functions, accuracy may be increased optimally by increasing L in a way that is dependent on the order of the quadrature rule (ℓ) and the growth of the integrand at ∞ . However, if the integrand is meromorphic, increasing L will also draw singularities toward the sampling domain, and accuracy will eventually degrade. Choosing L optimally to balance these two effects is nontrivial, and Boyd [12] offers concrete strategies and an insightful discussion. To illustrate the effect of different choices for L , we used (3.4) to compute the \mathcal{H}_2 norm of the Heat Model from the NICONET Benchmark collection [14]. With only 20 function evaluations and using $L = 0.486$, we approximated $\|H\|_{\mathcal{H}_2}$ with a relative error of $2.8 \cdot 10^{-7}$. Even using only 10 function evaluations (while keeping the same L value) resulted in a relative error of $2.2 \cdot 10^{-4}$. When one considers that the usual computational task involved in computing the \mathcal{H}_2 norm involves the solution of a (large) Lyapunov equation, the ability to compute the \mathcal{H}_2 norm to such great accuracy with only 10 function evaluations suggests the power that effective numerical quadrature can bring. Note that in this example, the function behaves quite well. If the function has many nearly unstable poles,

then determining an optimal L will not be as simple. To provide some contrast, if we decrease L to $L = 0.1$ then with 20 function evaluations, the \mathcal{H}_2 norm is estimated with a worse relative error of $2.7 \cdot 10^{-4}$. Likewise, if we increase L to $L = 1$ then we also obtain a degraded relative error of $7.8 \cdot 10^{-4}$. The price of a poor choice of L may be a significant increase in quadrature order so as to compensate for the loss of accuracy: If we choose an even smaller L value such as $L = 0.01$, 60 function evaluations will give a relative error of $4.9 \cdot 10^{-3}$, and increasing the number of function evaluations to 90 recovers an accuracy of $8.4 \cdot 10^{-4}$. We do not discuss the interesting and important question of how best to choose L further here, since we have introduced this quadrature rule here only to illustrate our approach.

We now adapt the B/CC quadrature rule in order to modify the objective function for VF. In the k th step, the r th order rational approximant is defined as before: $H_r^{(k)}(s) = \frac{\sum_{j=1}^r \frac{\phi_j^{(k)}}{s - \lambda_j^{(k)}}}{1 + \sum_{j=1}^r \frac{\varphi_j^{(k)}}{s - \lambda_j^{(k)}}}$. The poles $\lambda^{(k+1)}$ are determined from the r roots of $1 + \sum_{j=1}^r \frac{\varphi_j^{(k)}}{s - \lambda_j^{(k)}} = 0$. Now, $\phi_j^{(k)}$ and $\varphi_j^{(k)}$, will be determined from the solution of the successive weighted least squares problems

$$\|\Delta \left(\mathcal{A}(\lambda^{(k+1)})x^{(k+1)} - h \right)\|_2 \rightarrow \min, \quad k = 0, 1, 2, \dots, \quad (3.6)$$

where $x^{(k+1)} = (\phi_1^{(k+1)} \phi_2^{(k+1)} \dots \phi_r^{(k+1)} \varphi_1^{(k+1)} \varphi_2^{(k+1)} \dots \varphi_r^{(k+1)})^T$,

$$\mathcal{A}(\lambda) = \begin{pmatrix} \frac{1}{\xi_1 - \lambda_1} & \frac{1}{\xi_1 - \lambda_2} & \dots & \frac{1}{\xi_1 - \lambda_r} & \frac{-H(\xi_1)}{\xi_1 - \lambda_1} & \frac{-H(\xi_1)}{\xi_1 - \lambda_2} & \dots & \frac{-H(\xi_1)}{\xi_1 - \lambda_r} \\ \frac{1}{\xi_2 - \lambda_1} & \frac{1}{\xi_2 - \lambda_2} & \dots & \frac{1}{\xi_2 - \lambda_r} & \frac{-H(\xi_2)}{\xi_2 - \lambda_1} & \frac{-H(\xi_2)}{\xi_2 - \lambda_2} & \dots & \frac{-H(\xi_2)}{\xi_2 - \lambda_r} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \frac{1}{\xi_{\ell-1} - \lambda_1} & \frac{1}{\xi_{\ell-1} - \lambda_2} & \dots & \frac{1}{\xi_{\ell-1} - \lambda_r} & \frac{-H(\xi_{\ell-1})}{\xi_{\ell-1} - \lambda_1} & \frac{-H(\xi_{\ell-1})}{\xi_{\ell-1} - \lambda_2} & \dots & \frac{-H(\xi_{\ell-1})}{\xi_{\ell-1} - \lambda_r} \\ \frac{1}{\xi_{\ell} - \lambda_1} & \frac{1}{\xi_{\ell} - \lambda_2} & \dots & \frac{1}{\xi_{\ell} - \lambda_r} & \frac{-H(\xi_{\ell})}{\xi_{\ell} - \lambda_1} & \frac{-H(\xi_{\ell})}{\xi_{\ell} - \lambda_2} & \dots & \frac{-H(\xi_{\ell})}{\xi_{\ell} - \lambda_r} \\ 1 & 1 & \dots & 1 & 0 & 0 & \dots & 0 \end{pmatrix}, \quad h = \begin{pmatrix} H(\xi_1) \\ H(\xi_2) \\ \vdots \\ H(\xi_{\ell-1}) \\ H(\xi_{\ell}) \\ M_+[H] \end{pmatrix},$$

and $\Delta = \text{diag}(\rho_1, \rho_2, \dots, \rho_{\ell}, \rho_+)$ with nodes $\xi_j = iL \cot\left(\frac{j\pi}{\ell+1}\right)$

and weights $\rho_j = \csc\left(\frac{j\pi}{\ell+1}\right) \sqrt{\frac{L\pi}{(\ell+1)}}$ for $j = 1, \dots, \ell$ and $\rho_+ = \sqrt{\frac{\pi}{L(\ell+1)}}$ (3.7)

determined by the quadrature rule (3.4). This describes the main iteration of our quadrature-based variant of VF. We will refer to this variant as **QuadVF**. The term M_+ from (3.5) is retained and given double weight, since $M_+[H] = M_-[H]$ for real systems. Notice that the weighting matrix Δ is fixed with respect to k and that the quadrature nodes are closed under conjugation: $\xi_j = \overline{\xi_{\ell+1-j}}$, halving the number of function evaluations needed to implement the formula. This symmetry is also reflected in the weights: $\rho_j = \rho_{\ell+1-j}$.

3.4. Numerical Comparisons.

3.4.1. Heat Model: VF vs. QuadVF. We use the aforementioned Heat Model for this example. We take $\ell = 20$ samples (requiring only 10 function evaluations due to the complex conjugate sampling points) and apply both VF and QuadVF to construct order $r = 4$ rational approximants. In this case, the sampling nodes for both VF and QuadVF nodes are contained in $i[2.7705 \times 10^{-2}, 2.4527]$; only the distribution of the nodes is different. The resulting relative \mathcal{H}_2

error norms are 8.4776×10^{-1} for VF and 6.9326×10^{-3} for QuadVF. The numbers for the relative \mathcal{H}_∞ error norms were even more revealing: 1.6392 for VF and 6.7765×10^{-4} for the QuadVF.

Note that the poor approximation resulting from VF is not due to a large residual for the underlying LS problem. On the contrary, VF leads to a relative LS residual norm of 1.8943×10^{-3} , representing a very accurate solution to the discrete LS problem; for QuadVF, the relative residual norm is 3.5430×10^{-4} , yielding in this case not only an accurate solution to the discrete LS problem but also a comparable level of accuracy as an ideal \mathcal{H}_2 -optimal reduced model of the same order. VF does a great job in minimizing the least-squares error over the given samples; however the samples are local in nature and do not reflect the global \mathcal{H}_2 and/or \mathcal{H}_∞ behavior. By choosing the sampling nodes from an appropriate quadrature rule, the discrete error that is minimized becomes a much better approximation to the true \mathcal{H}_2 error, leading ultimately to a better rational approximation.

3.4.2. FOM Model: VF vs QuadVF. We repeat the same numerical experiments for the FOM Model by taking $\ell = 50$ samples (requiring only 25 function evaluations) and applying VF and QuadVF as before. For this model, we construct an order $r = 12$ rational approximant. The sampling interval for VF and QuadVF is the same: $\xi \in i[3.0810, 1.6213 \times 10^3]$, again differing only by their distribution in the interval. The resulting relative \mathcal{H}_2 error norms are: 3.0903×10^{-2} for VF, and 1.8561×10^{-3} for QuadVF; QuadVF outperforms VF by more than an order of magnitude in terms of accuracy. Similar results are found for \mathcal{H}_∞ performance as well with VF and QuadVF leading to relative \mathcal{H}_∞ error norms of, respectively, 7.6430×10^{-2} and 3.2204×10^{-3} . As in the previous example, the difference in the approximation quality is not due to the underlying discrete LS residuals. Both VF and QuadVF produced very accurate LS solutions with relative residual norms of 8.1809×10^{-5} and 4.6997×10^{-5} , respectively. The improved node and weight selection of QuadVF appears to be the determining factor for the improved quality of the rational approximation. However, even QuadVF does not match the high-fidelity optimal rational approximations. For this example, IRKA produces final reduced models with relative \mathcal{H}_2 and \mathcal{H}_∞ errors of 1.9200×10^{-4} and 2.1157×10^{-4} , respectively; an order of magnitude better in both cases.

3.4.3. Heat Model: QuadVF vs IRKA. QuadVF is based on the discretization of the true \mathcal{H}_2 norm. Therefore in this example, we investigate numerically how the solution of the quadrature-based discrete \mathcal{H}_2 minimization problem compares to the the solution of the continuous \mathcal{H}_2 problem by IRKA as the number of sampling points ℓ increases. We use the Heat Model and construct order $r = 2$ rational approximants using QuadVF and IRKA. Let H , H_1, H_2 denote, respectively, the full-order model, the reduced model by IRKA and the reduced model by QuadVF. In Table 3.4.3 below, we list the relative \mathcal{H}_2 distances between H_1 and H_2 as ℓ increases in addition to the relative \mathcal{H}_2 distances between the full and two reduced models:

ℓ	$\frac{\ H_1 - H_2\ _{\mathcal{H}_2}}{\ H_1\ _{\mathcal{H}_2}}$	$\frac{\ H - H_1\ _{\mathcal{H}_2}}{\ H\ _{\mathcal{H}_2}}$	$\frac{\ H - H_2\ _{\mathcal{H}_2}}{\ H\ _{\mathcal{H}_2}}$
10	1.1919×10^{-2}	3.9483×10^{-2}	4.1348×10^{-2}
100	3.8795×10^{-3}	3.9483×10^{-2}	3.9681×10^{-2}
1000	1.0239×10^{-3}	3.9483×10^{-2}	3.9497×10^{-2}
5000	5.2313×10^{-4}	3.9483×10^{-2}	3.9487×10^{-2}
15000	4.4926×10^{-4}	3.9483×10^{-2}	3.9486×10^{-2}

TABLE 3.1
Relative \mathcal{H}_2 distances vs ℓ

Table 3.4.3 illustrates that for this numerical example, as ℓ increases, the solution of the discrete \mathcal{H}_2 problem via QuadVF is converging to the true \mathcal{H}_2 solution. This is an encouraging result confirming that an effective quadrature-based selection for the discretized \mathcal{H}_2 problem might yield rational

approximants close to those of the true, continuous problem. These issues will be further studied and presented in [7]. For comparison, we increased the sampling size for the VF as well. However, even with $\ell = 15000$, VF produced a rational approximant, $H_3(s)$, with relative \mathcal{H}_2 distances

$$\frac{\|H_1 - H_3\|_{\mathcal{H}_2}}{\|H_1\|_{\mathcal{H}_2}} = 9.8470 \times 10^{-1} \quad \text{and} \quad \frac{\|H - H_3\|_{\mathcal{H}_2}}{\|H_3\|_{\mathcal{H}_2}} = 9.8503 \times 10^{-1},$$

The contrast with QuadVF underscores the value of sampling guided by an effective quadrature rule.

3.4.4. ISS1R Module: QuadVF vs VF. We use the ISS 1R module [26] with $n = 270$ and approximate it with a model of order $r = 16$. We first use QuadVF and 25 function evaluations ($\ell = 50$ nodes in 25 complex conjugate pairs). The relative \mathcal{H}_2 and \mathcal{H}_∞ errors of QuadVF were 7.2156×10^{-2} and 2.4448×10^{-2} . The relative \mathcal{H}_2 and \mathcal{H}_∞ errors of IRKA (using the same initial poles as QuadVF) were, respectively, 1.4474×10^{-2} and 5.5595×10^{-3} – lower, as expected. Next, for comparisons, we use the same interval $\mathbb{H}[1.2324 \times 10^{-1}, 6.4853 \times 10^1]$ containing the quadrature nodes, and replace the nodes by the same number of (i) linearly spaced points, and (ii) logarithmically spaced points. Then VF is run with those points. For the case of linearly spaced points, VF produced relative \mathcal{H}_2 and \mathcal{H}_∞ errors of 104.24% and 99.79%, respectively, almost two orders of magnitude higher errors than QuadVF. For logarithmically spaced points, VF performed better and produced relative \mathcal{H}_2 and \mathcal{H}_∞ errors 3.2872×10^{-1} and 1.2257×10^{-1} ; still much less accurate than QuadVF. The Bode plots of the full-model and all four rational approximants are shown in Figure 3.1.

REMARK 3.1. Recently, Hochman, Leviatan and White [33] also formulated rational least squares approximation using the information from the quadrature nodes. There, the problem is to find real valued potential U that satisfies Laplace equation in a simply connected domain $\Omega \subset \mathbb{R}^2$ and the Dirichlet boundary condition $U|_\Gamma = f$ on the boundary curve Γ of Ω . The idea is to approximate U with the truncated real part \hat{U} of a weighted sum W of complex dipole potentials, and to enforce the boundary condition on Γ by minimizing $\|\hat{U} - f\|_\Gamma$, where $\|\cdot\|_\Gamma$ is induced by the inner product $(u, v)_\Gamma = \int_0^1 u(z(s))v^*(z(s))\lambda(s)ds$ along Γ . (Here $z(s)$ is a parametrization of Γ and $\lambda(s)$ is a positive weight function.) Discretizing the norm introduces the quadrature nodes.

3.5. Vector fitting in a discrete Sobolev norm. Incorporating derivative information into function approximation strategies (e.g., by penalizing roughness of the error function, or forcing Hermite interpolation at selected points) often can produce significantly higher fidelity approximations at only marginally increased cost. Many interpolatory model reduction methods, including IRKA, construct rational approximants, $H_r(s)$, that match the value of $H(s)$ together with some of its derivatives at selected interpolation points, a type of generalized Hermite interpolation. Since derivatives in the frequency domain are associated with moments in the time domain, the expression “moment matching methods”, as exemplified e.g., by the “Padé via Lanczos” (PVL) method [23], refers also to a similar generalized Hermite interpolation strategy.

Chen, Zheng, and Fang [15] included derivatives in their modification of VF, leading to what they termed “Moment Matching Vector Fitting”, a multipoint moment matching scheme with the approximating rational function given in barycentric form. Derivative conditions that are compatible with the VF framework can be obtained by differentiating the expression, $H(s)d(s) = n(s)$. For example, to match the first derivative, one uses the condition $d'(s)H(s) + d(s)H'(s) - n'(s) = 0$, which is a linear expression in the coefficients of $n(s)$ and $d(s)$. Based on this expression and similar ones for higher derivatives, Chen, *et al.* in [15] derived a system of equations that incorporate derivative conditions. The assumed barycentric form of the approximant then produces a coefficient matrix with a Cauchy-like structure similar to what is obtained for VF.

In this section, we develop a somewhat different approach toward incorporating derivative information into VF. Analogous to our approach for QuadVF, we begin with an approximation problem

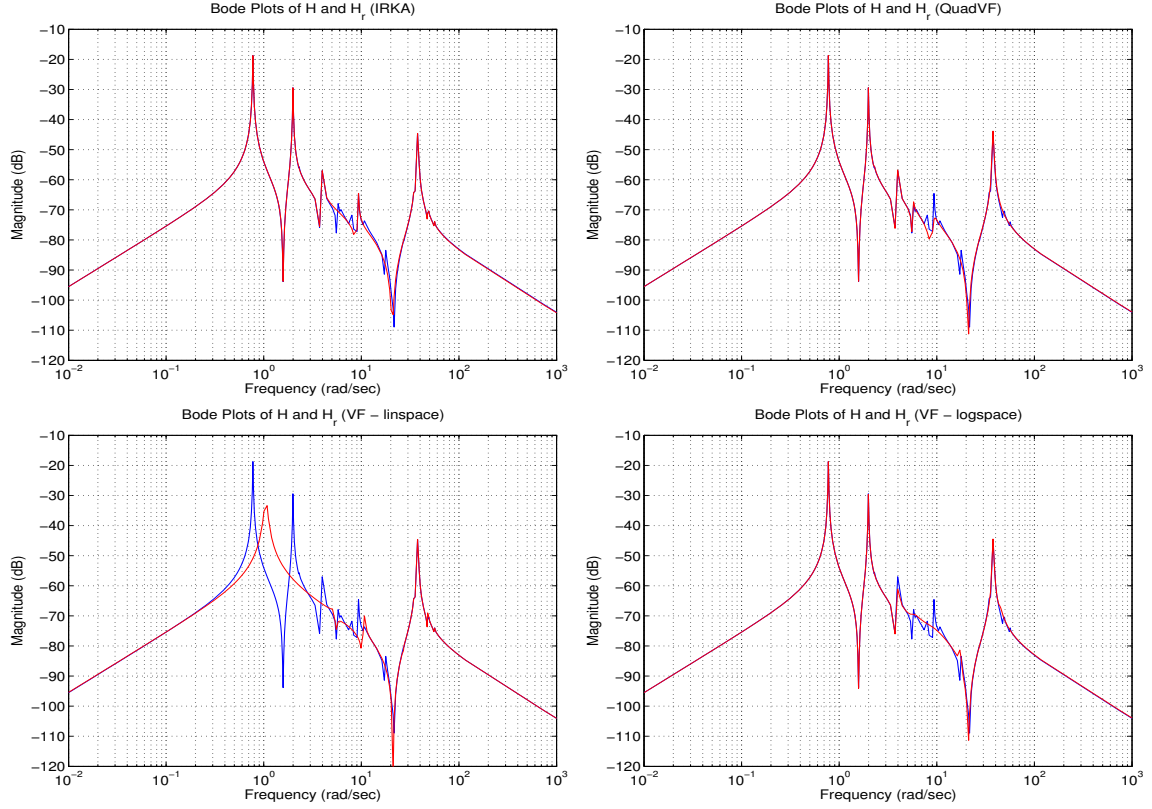


FIG. 3.1. Amplitude Bode plots of the original system (blue line in every plot) and four rational approximations (red line) (Top-left: QuadVF, Top-right: IRKA, Bottom-left: VF with linearly spaced points, Bottom-right: VF with logarithmically spaced points.)

formulated with respect to an appropriate continuous norm and then discretize, making use of effective quadrature points and weights. Derivative conditions arise differently than in [15], leading to a significant difference in the diagonal scaling.

Given $H(s)$ and sampling nodes, ξ_i , we seek a rational function, $H_r(s)$, that will yield good approximations not only to $H(\xi_i)$ but also to $H'(\xi_i)$, in the least-squares sense. Restated formally, the problem is to find an r th order stable rational approximant:

$$H_r(s) = \frac{n(s)}{d(s)} \equiv \frac{\sum_{j=1}^r \frac{\phi_j}{s - \lambda_j}}{\sum_{j=1}^r \frac{\varphi_j}{s - \lambda_j} + 1}, \quad (3.8)$$

$$\text{such that } \sum_{i=1}^{\ell} (\rho_{i0}^2 |H_r(\xi_i) - H(\xi_i)|^2 + \rho_{i1}^2 |H'_r(\xi_i) - H'(\xi_i)|^2) \longrightarrow \min.$$

There is a significant difference in our problem formulation (3.8) and that of [15]. We view the minimization problem considered in (3.8), as the discretization of a minimization problem formulated now with respect to a continuous Sobolev-type \mathcal{H}_2 norm,

$$\|H - H_r\|_{\star}^2 = \|H - H_r\|_{\mathcal{H}_2}^2 + \|H' - H'_r\|_{\mathcal{H}_2}^2,$$

and apply an appropriate quadrature rule (see e.g. [36]) to determine nodes ξ_i and weights ρ_{i0} , ρ_{i1} in (3.8). This has the effect of penalizing roughness of the error function, $H - H_r$, and will

yield a different rational approximant to $H(s)$. For an overview of derivative-weighted least squares approximation, we refer to [24, §3.2.3].

To arrive at a VF iteration for (3.8), first approximate the derivative error

$$\begin{aligned} H'(s) - H'_r(s) &= H'(s) + \frac{n(s)d'(s) - n'(s)d(s)}{d^2(s)} = \frac{d(s)H'(s) + d'(s)H_r(s) - n'(s)}{d(s)} \\ &\approx \frac{d(s)H'(s) + d'(s)H(s) - n'(s)}{d(s)}. \end{aligned} \quad (3.9)$$

Then approximating the \mathcal{H}_2 norms with quadrature rules and incorporating the rescaling characteristic of the SK iteration produces a weighted LS problem that appears as

$$\begin{aligned} \|H - H_r\|_*^2 &\approx \sum_{i=1}^{\ell} \frac{\rho_{i0}^2}{|d^{(k)}(\xi_i)|^2} \left| \sum_{j=1}^r \frac{\phi_j^{(k+1)}}{\xi_i - \lambda_j^{(k)}} - \sum_{j=1}^r \frac{H(\xi_i)}{\xi_i - \lambda_j^{(k)}} \varphi_j^{(k+1)} - H(\xi_i) \right|^2 \\ &\quad + \sum_{i=1}^{\ell} \frac{\rho_{i1}^2}{|d^{(k)}(\xi_i)|^2} \left| \sum_{j=1}^r \frac{-\phi_j^{(k+1)}}{(\xi_i - \lambda_j^{(k)})^2} + \sum_{j=1}^r \left(\frac{H(\xi_i)}{(\xi_i - \lambda_j^{(k)})^2} - \frac{H'(\xi_i)}{\xi_i - \lambda_j^{(k)}} \right) \varphi_j^{(k+1)} - H'(\xi_i) \right|^2. \end{aligned}$$

The structure of the LS matrix (cf.(2.6)-(2.7)) becomes more complicated: Set $D'_\xi = \text{diag}(h')$, $h' = (H'(\xi_i))_{i=1}^\ell$, $W_j = \text{diag}(\rho_{ij})_{i=1}^\ell$, ($j = 0, 1$), $\Delta^{(k)} = \text{diag}(1/|d^{(k)}(\xi_i)|)_{i=1}^\ell$, and $\mathcal{C}_{ij}^{(k)} = 1/(\xi_i - \lambda_j^{(k)})$. The new LS problem reads

$$\left\| \begin{pmatrix} W_0 \Delta^{(k)} & 0 \\ 0 & W_1 \Delta^{(k)} \end{pmatrix} \left\{ \begin{pmatrix} \mathcal{C}^{(k)} & -D_\xi \mathcal{C}^{(k)} \\ -(\mathcal{C}^{(k)} \circ \mathcal{C}^{(k)}) & D_\xi(\mathcal{C}^{(k)} \circ \mathcal{C}^{(k)}) - D'_\xi \mathcal{C}^{(k)} \end{pmatrix} \begin{pmatrix} \phi_{1:r}^{(k+1)} \\ \varphi_{1:r}^{(k+1)} \end{pmatrix} - \begin{pmatrix} h \\ h' \end{pmatrix} \right\} \right\|_2 \rightarrow \min \quad (3.10)$$

where “ \circ ” denotes the Hadamard matrix product.

The final expression of (3.9) is approximate because a correction term, $\frac{d'(s)}{d(s)}(H(s) - H_r(s))$, has been dropped. This additional term may be retained and incorporated into the final LS problem (3.10), although the additional complexity might not be justified. For example, one may approximate the correction term evaluated at $s = \xi_i$ as

$$\frac{d'(\xi_i)}{d(\xi_i)}(H(\xi_i) - H_r(\xi_i)) \approx \frac{d^{(k+1)}(\xi_i)}{d^{(k)}(\xi_i)} \left(H(\xi_i) - \frac{n^{(k)}(\xi_i)}{d^{(k)}(\xi_i)} \right).$$

This yields a more complicated, though similarly structured LS coefficient matrix. We believe that this is not necessary in practice since the effect of penalizing derivative error appears to be achieved quite effectively with the simpler expression. Note that the first part of the Sobolev error expression, $\|H - H_r\|_*^2$, penalizes the magnitude of $H(\xi_i) - H_r(\xi_i)$ suggesting that the correction term that has been omitted will become small in any case. In addition, as the iteration progresses, the residues of $d(s)$ are expected to converge to 0, so that $d(s) \rightarrow 1$ and $d'(s) \rightarrow 0$ almost everywhere, further diminishing the term that has been omitted.

Adopting the pole relocation and rescaling strategies characteristic of VF, we find

PROPOSITION 3.1. *By a change of barycentric representation, the LS problem (3.10) can be replaced by*

$$\left\| \begin{pmatrix} W_0 & 0 \\ 0 & W_1 \end{pmatrix} \left\{ \begin{pmatrix} \mathcal{C}^{(k+1)} & -D_\xi \mathcal{C}^{(k+1)} \\ -(\mathcal{C}^{(k+1)} \circ \mathcal{C}^{(k+1)}) & D_\xi(\mathcal{C}^{(k+1)} \circ \mathcal{C}^{(k+1)}) - D'_\xi \mathcal{C}^{(k+1)} \end{pmatrix} \begin{pmatrix} \tilde{\phi}_{1:r}^{(k+1)} \\ \tilde{\varphi}_{1:r}^{(k+1)} \end{pmatrix} - \begin{pmatrix} h \\ h' \end{pmatrix} \right\} \right\|_2 \rightarrow \min, \quad (3.11)$$

where $\mathcal{C}_{ij}^{(k+1)} = 1/(\xi_i - \lambda_j^{(k+1)})$, and $(\lambda_j^{(k+1)})_{j=1}^\ell$ are the zeros of $d^{(k)}(s)$.

Proof. Consider all iterations done up through step $k + 1$ to have been done with fixed poles, namely $\lambda_j^{(k+1)}$ for $j = 1, \dots, \ell$. If we want the next iterate to be represented in the barycentric form with the nodes $\lambda_j^{(k+1)}$, then, to be consistent with the definition of the iterations (2.1), the scaling factors $1/|d^{(k)}(\xi_i)|$ must be computed using the barycentric form of $H_r^{(k)} = n^{(k)}/d^{(k)}$ based on the nodes $\lambda_j^{(k+1)}$. Now, if we represent $n^{(k)}/d^{(k)}$, with $d^{(k)}$ as in (2.9), with the nodes $\lambda_j^{(k+1)}$, then we

obtain $\frac{n^{(k)}(s)}{d^{(k)}(s)} = \frac{\sum_{j=1}^r \frac{\hat{\phi}_j^{(k+1)}}{s - \lambda_j^{(k+1)}}}{1}$. Hence, in this representation the scaling factors are 1. \square

The Sobolev norm-based VF iteration described in Proposition 3.1 will be called **SobVF** and will be run typically until the nodes $\lambda_j^{(k)}$ converge (numerically) at some index k_* . To compute our final rational approximant, we take the converged $\lambda_j^{(k_*)}$'s as the poles and solve LS problem

$$\left\| \begin{pmatrix} W_0 & 0 \\ 0 & W_1 \end{pmatrix} \left\{ \begin{pmatrix} \mathcal{C}^{(k_*)} \\ -(\mathcal{C}^{(k_*)} \circ \mathcal{C}^{(k_*)}) \end{pmatrix} \phi_{1:r} - \begin{pmatrix} h \\ h' \end{pmatrix} \right\} \right\|_2 \rightarrow \min \quad (3.12)$$

the compute the final residues ϕ_j .

REMARK 3.2. Even though obtaining the derivative information may not be always feasible (e.g., in the data driven setting), in many cases $H'(s)$ can be computed without much additional cost. For example, if a state space representation $H(s) = \mathbf{C}(s\mathbf{I} - \mathbf{F})^{-1}\mathbf{B}$ is available, then computing $H'(s) = -\mathbf{C}(s\mathbf{I} - \mathbf{F})^{-2}\mathbf{B}$ is not expensive if the function evaluation is performed using, for example, sparse direct solvers or a Hessenberg decomposition-based method for dense computations [6]. The evaluation of $H(s)$ already requires the computation of a decomposition of $(s\mathbf{I} - \mathbf{F})$ at the node $s = \xi_i$. Since evaluating $H'(s)$ at the node $s = \xi_i$ requires solving a linear system with the same coefficient matrix, the triangular factors can be reused, and $H(s)$ and $H'(s)$ at the node ξ_i are obtained with only small additional cost.

3.5.1. Numerical Examples for SobVF. We illustrate the effectiveness of SobVF using two models from the NICONET Benchmark Collection, comparing results with VF. Since SobVF uses both $H(s)$ and $H'(s)$ at the sampling nodes, we use twice the number of nodes in VF in order to present a fair comparison for VF; that is, if we use ℓ nodes in (3.8), we will employ 2ℓ in VF. For brevity, instead of adapting and giving details of a Hermite quadrature rule, we simply use the weights and the nodes of the Clenshaw-Curtis formula from §3.3 in both examples.

EXAMPLE 3.2. The first example is the Building Model from the NICONET benchmark collection with order $n = 48$. We have chosen this model since it is very hard to approximate and a high-fidelity approximation is achieved only for large r values [5]. For example, to reach a relative \mathcal{H}_2 error norm of 10^{-4} , even the optimal rational approximation method IRKA requires $r = 40$ and then yields a relative \mathcal{H}_2 of 1.18×10^{-4} . We pick $r = 40$ and obtain the nodes and weights using §3.3. The range of nodes for VF and SobVF is the same; only the distribution is different. For $\ell = 25$, VF using $2\ell = 50$ logarithmically spaced nodes yields a relative \mathcal{H}_2 error norm of 1.564 – quite a poor approximation. On the other hand, using SobVF as in (3.8) with $\ell = 25$ nodes yields a rational approximant with a relative \mathcal{H}_2 error of $6.56 \cdot 10^{-3}$. This constitutes a three order-of-magnitude improvement over what VF provides without greater computational cost; recall VF used twice the number of nodes as SobVF.

EXAMPLE 3.3. We consider the Beam Model for the NICONET benchmark collection. This model has order $n = 348$. Using $\ell = 25$ as in the previous example for SobVF approximation and $2\ell = 50$ nodes for VF approximation, we obtain relative \mathcal{H}_2 errors of 1.29 for VF and 0.16 for SobVF. To obtain better approximants, we double the number of nodes to $\ell = 50$, leading to a relative \mathcal{H}_2 error norm of $4.84 \cdot 10^{-2}$ for VF and $2.85 \cdot 10^{-4}$ for SobVF. We observe that for $r = 40$, the optimal approximation method IRKA yield a relative error of $2.09 \cdot 10^{-4}$. So, using $\ell = 50$ nodes,

SobVF very nearly achieves the accuracy captures the accuracy of a locally optimal approximant. To investigate how the approximants change, we increase the order to $r = 70$. Curiously, this caused a *higher* relative error of $1.84 \cdot 10^{-1}$ for VF. This is mainly due to the numerical ill-conditioning of the underlying LS problem induced by increasing r . These issues are explained in more detail in §4.3. On the other hand, increasing r to 70 had no apparent adverse effect on the SobVF; the relative error decreased to $4.17 \cdot 10^{-6}$. For comparison, note that for $r = 70$, the relative \mathcal{H}_2 error produced by IRKA is $5.10 \cdot 10^{-7}$. Although IRKA is still better (as expected), the SobVF approximation is achieving close to the same accuracy.

In both of the experiments described above, the SobVF approximation was substantially more accurate than a QuadVF approximation produced with the same set of nodes and weights. As previously stated, this will not even be the best performance that can be expected from SobVF. The full-potential of (3.8) will be realized once we adopt an appropriate quadrature rule, much as we did in §3.3 to produce QuadVF. We defer these considerations to a later time.

4. Practical Issues. We focus on the convergence behavior and some practical issues impacting the numerical implementation of both VF and QuadVF.

4.1. Unstable nodes mirroring and scaling. One of the advantages of the pole relocation step in VF is that the emergence of unstable poles can be resolved and the iterates can be steered to a stable approximant. This is achieved by reflecting those unstable nodes (poles) that are in \mathbb{C}_+ with respect to the imaginary axis and placing them in \mathbb{C}_- . The same procedure is also employed in IRKA. Let $\tilde{n}^{(k)}(s)/\tilde{d}^{(k)}(s)$ be the current approximation, $\lambda_j^{(k+1)}$ denote the the originally computed set of zeros of $\tilde{d}^{(k)}$ and $\lambda_{j_t}^{(k+1)}$, $t = 1, \dots, p$, be the $p < r$ of these poles that are in \mathbb{C}_+ . Then, VF replaces $\lambda_{j_t}^{(k+1)}$ with $-\lambda_{j_t}^{(k+1)}$ while keeping the remaining stable ones as is to obtain the new set of poles, to be denoted by $\hat{\lambda}_j^{(k+1)}$ with $\hat{\lambda}_t^{(k+1)} = -\lambda_{j_t}^{(k+1)}$, $t = 1, \dots, p$. From a systems theoretic perspective, the mirroring of an unstable pole $\lambda_{j_t}^{(k+1)}$ corresponds to applying an all-pass filter $\Phi_{j_t}(s) = (s - \lambda_{j_t}^{(k+1)})/(s + \overline{\lambda_{j_t}^{(k+1)}})$ that changes the phase of the approximant, see [31]. Let $\hat{n}^{(k)}/\hat{d}^{(k)}$ be the barycentric representation corresponding to the nodes $\hat{\lambda}_j^{(k+1)}$. Then, VF proceeds by solving the LS problem $\|\mathcal{A}(\hat{\lambda}^{(k+1)})\hat{x}^{(k+1)} - h\|_2 \rightarrow \min$, instead of $\|\hat{\Delta}^{(k)}(\mathcal{A}(\hat{\lambda}^{(k+1)})\hat{x}^{(k+1)} - h)\|_2 \rightarrow \min$. This is not formally correct – since the poles are changed by an external intervention, pole relocation does not compensate diagonal scaling.

To make this step formally correct and interpretable in the framework of numerical linear algebra, we need the barycentric representation $\hat{n}^{(k)}(s)/\hat{d}^{(k)}(s)$ of $\tilde{n}^{(k)}(s)/\tilde{d}^{(k)}(s)$, and the corresponding diagonal scaling $\hat{\Delta}^{(k)} = \text{diag}(1/|\hat{d}^{(k)}(\xi_i)|)_{i=1}^\ell$ expressed using the new poles $\hat{\lambda}_j^{(k+1)}$ (cf. the proof of Proposition 3.1). Such a representation can be directly written down using

$$\frac{\tilde{n}^{(k)}(s)}{\tilde{d}^{(k)}(s)} \equiv \frac{\hat{n}^{(k)}(s)}{\hat{d}^{(k)}(s)} = \frac{\sum_{j=1}^r \frac{\alpha_j^{(k)}}{s - \hat{\lambda}_j^{(k+1)}}}{\sum_{j=1}^p \frac{\beta_j^{(k)}}{s - \hat{\lambda}_j^{(k+1)}} + 1}, \quad \hat{d}^{(k)}(s) = \sum_{j=1}^p \frac{\beta_j^{(k)}}{s - \hat{\lambda}_j^{(k+1)}} + 1, \quad (4.1)$$

where the $\beta_j^{(k)}$'s must be determined so that the zeros of $\hat{d}^{(k)}(s)$ are $\lambda_{j_t}^{(k+1)}$, $t = 1, \dots, p$. This is an eigenvalue assignment problem in disguise and we use [42] to get

$$\beta_j^{(k)} = \frac{\prod_{\ell=1}^p (\hat{\lambda}_j^{(k+1)} + \hat{\lambda}_\ell^{(k+1)})}{\prod_{\ell=1, \ell \neq j}^p (\hat{\lambda}_j^{(k+1)} - \hat{\lambda}_\ell^{(k+1)})}, \quad j = 1, \dots, p. \quad (4.2)$$

PROPOSITION 4.1. Let $\widehat{d}^{(k)}(s)$ be defined as in (4.1), (4.2). Then for any $\omega \in \mathbb{R}$, $|\widehat{d}^{(k)}(i\omega)| = 1$ and the diagonal scaling matrix $\widehat{\Delta}^{(k)}$ is unitary; the solution does not change from that of the unscaled problem.

Proof. Note that

$$\prod_{j=1}^p (i\omega - \hat{\lambda}_j^{(k+1)}) \widehat{d}^{(k)}(i\omega) = \prod_{j=1}^p (i\omega - \hat{\lambda}_j^{(k+1)}) \left(\sum_{j=1}^p \frac{\beta_j^{(k)}}{i\omega - \hat{\lambda}_j^{(k+1)}} + 1 \right) = \prod_{j=1}^p (i\omega + \hat{\lambda}_j^{(k+1)}).$$

Recall that the $\hat{\lambda}_j^{(k+1)}$, $j = 1, \dots, p$, are closed under complex conjugation. The claim follows. \square

Proposition 4.1 justifies proceeding with the same VF scheme after mirroring unstable poles, as if nothing had happened. The same applies to the SobVF approximation described in §3.5.

4.2. Numerical convergence and stopping criterion. A theoretical convergence analysis of VF that determines conditions on $H(s)$ and the sampling nodes so as to guarantee convergence of VF remains an open problem. An instructive analysis by Lefteriu and Antoulas [39] showed (using a synthetic example with $r = 2$) that the fixed points of the VF iterations can actually be repellant and so that the iteration may diverge. Convergence behavior in realistic, large-scale settings appears not yet to have been analyzed, and, to the best of our knowledge, there are no published stopping criteria for the VF iteration that can be justified rigorously by a rigorous error or perturbation analysis. In this section, we try to shed some light on these issues.

Assume now the setting of §2.2 with an ideal convergence scenario: Suppose that for some index k , the zeros and the poles of $\widehat{d}^{(k)}(s)$ can be *numerically matched*, so that $\lambda_j^{(k+1)} \approx \lambda_j^{(k)}$, and hence $\widehat{d}^{(k)}(s) \cong 1$. Restated, this means that the optimal matching distance

$$\Omega_k = \min_{\sigma \in \mathbb{S}_r} \max_{j=1:r} |\lambda_j^{(k)} - \lambda_{\sigma(j)}^{(k+1)}| \quad (\text{here } \mathbb{S}_r \text{ denotes the permutation group}) \quad (4.3)$$

between $(\lambda_j^{(k+1)})_{j=1}^r$ and $(\lambda_j^{(k)})_{j=1}^r$ as well as $\max_j |\widetilde{\varphi}_j^{(k)}|$ are all *sufficiently small*. The important tasks that arise here are determining k and quantifying and justifying how small is “*sufficiently small*”? The following observations provide the key insights.

(i) Recall that $\boldsymbol{\lambda}^{(k+1)}$ is the spectrum of $\text{diag}(\boldsymbol{\lambda}^{(k)}) + \widetilde{\boldsymbol{\varphi}}^{(k)} \mathbf{e}^T$, and thus can be considered as the spectrum of a rank-one perturbation of the matrix $\text{diag}(\boldsymbol{\lambda}^{(k)})$. Hence, by [11, Exercise VIII.3.2],

$$\Omega_k \leq (2r - 1) \|\widetilde{\boldsymbol{\varphi}}^{(k)} \mathbf{e}^T\|_2 \leq \sqrt{r} (2r - 1) \|(\widetilde{\varphi}_j^{(k)})_{j=1}^r\|_2 \leq r(2r - 1) \max_j |\widetilde{\varphi}_j^{(k)}|, \quad (4.4)$$

where Ω_k is the optimal matching distance defined in (4.3). In other words, by monitoring $\widetilde{\boldsymbol{\varphi}}^{(k)}$, we can determine in advance when $\lambda_j^{(k)}$ converges (up to a predetermined tolerance) and thus end the pole identification phase.

(ii) Moreover, it can be checked that, with proper permutation matching used to enumerate $(\lambda_j^{(k+1)})_{j=1}^r$, the element-wise relative differences between $\mathcal{A}(\boldsymbol{\lambda}^{(k+1)})$ and $\mathcal{A}(\boldsymbol{\lambda}^{(k)})$ are bounded by

$$\max_{i,j} |(\mathcal{A}(\boldsymbol{\lambda}^{(k)}))_{ij} - \mathcal{A}(\boldsymbol{\lambda}^{(k+1)})_{ij}| / \mathcal{A}(\boldsymbol{\lambda}^{(k+1)})_{ij} \leq \frac{\Omega_k}{\mu_k}, \quad \text{where } \mu_k = \min_{i=1:\ell} \min_{j=1:r} |\xi_i - \lambda_j^{(k)}|. \quad (4.5)$$

Note that we can use (4.4) to estimate in advance that the difference (4.5) is less than given ϵ by checking if $\Omega_k \leq \mu_k \epsilon$, i.e., if $\max_j |\widetilde{\varphi}_j^{(k)}| \leq \mu_k \epsilon / (2r^2 - r)$.

(iii) Finally, another plausible and justifiable backward stable stopping criterion with a given tolerance threshold ε can be seen in (2.11) with $k \leftarrow k - 1$ as follows: From the estimate

$$\left| \sum_{j=1}^r \frac{\widetilde{\varphi}_j^{(k)}}{\xi_i - \lambda_j^{(k)}} \right| \leq \sqrt{r} \frac{\|(\widetilde{\varphi}_j^{(k)})_{j=1}^r\|_2}{\mu_k} \leq r \max_{j=1:r} |\widetilde{\varphi}_j^{(k)}| \frac{1}{\mu_k},$$

valid for all $i = 1, \dots, \ell$, where μ_k is as defined in (4.5), we conclude that if $\max_j |\tilde{\varphi}_j^{(k)}| \leq \varepsilon \mu_k / r$, the residue identification is simple because $\tilde{n}^{(k)}(s) = \sum_{j=1}^r \frac{\tilde{\phi}_j^{(k)}}{s - \lambda_j^{(k)}}$ can be taken as the final approximant in the pole-residue representation but now with a relative backward error of at most ε in the measurements $H(\xi_i)$. However, to be on the safe side, the common practice of VF is to use the “converged” poles and then solve the LS problem $\|\mathcal{A}^{(k)}(:, 1:r)(\phi_j^{(k)})_{j=1}^r - h\|_2 \rightarrow \min$ to determine the residues.

In practice, when the VF iterations converge, one observes that $\|(\tilde{\varphi}_j^{(k)})_{j=1}^r\|_2$ tends to zero and the estimate (4.4) reliably predicts the change in the nodes $(\lambda_j^{(k)})_{j=1}^r$ from step k to step $k+1$. However, if unstable nodes appear, they are mirrored as explained in §4.1 and one works with the $\hat{\lambda}_j^{(k+1)}$'s instead of the $\lambda_j^{(k+1)}$'s, which, in turn, means that (4.4) does not apply. In fact, it can happen that at each iteration until the very end, a subset of the poles need to be flipped to \mathbb{C}_- and neither the $\tilde{d}^{(k)}(s)$ converge to unity nor the nodes $\lambda_j^{(k)}$ settle as $k \rightarrow \infty$. That, however, does not necessarily mean that the approximation is hopelessly bad. The following example illustrates this fact.

EXAMPLE 4.1. We take the Beam model with $n = 348$ from the NICONET collection and obtain order $r = 17$ and $r = 18$ approximants using $\ell = 25$ conjugate pairs of logarithmically spaced nodes ξ_i . The VF convergence history shown in Figure 4.1 illustrates two phenomena. In the figure on the left with $r = 18$, the value of $\|(\tilde{\varphi}_j^{(k)})_{j=1}^r\|_2$ settles around 8.042 while the maximal relative change of the nodes drops down to the level of 10^{-13} . Thus, VF converges but with $\|(\tilde{\varphi}_j^{(k)})_{j=1}^r\|_2 \neq 0$. The relative \mathcal{H}_2 error norm of the resulting approximant is $5.11 \cdot 10^{-2}$. The right figure, on the other hand, with $r = 17$ shows a zigzag pattern for $\|(\tilde{\varphi}_j^{(k)})_{j=1}^r\|_2$ (indicating two accumulation points of the vectors $\tilde{\varphi}_j^{(k)}$, $k = 1, 2, \dots$) and the $\mathcal{O}(1)$ relative changes in the nodes from step to step (indicating here too two accumulation points, where the zigzag comes from computing the relative distances). Thus, neither $\{\tilde{\varphi}^{(k)}\}$ nor $\{\lambda^{(k)}\}$ converges. However, the iteration exhibits a periodicity in the behavior of the nodes. With the lag of 2 iterations, we see that $\|(\lambda^{(k)} - \lambda^{(k+2)})/\lambda^{(k)}\|_\infty$ drops to the level of 10^{-10} . In other words, the nodes cycle with the period of 2. The relative \mathcal{H}_2 error norms of the approximants are around $2.11 \cdot 10^{-2}$ and $2.45 \cdot 10^{-2}$, depending on the index k . It should be noted that the patterns shown on the right figure are not due to the flipping of unstable nodes. Even when that mechanism is switched off, in this example we observe nearly the same periodic behaviors but in that case with an eventual unstable approximant, resulting in infinite \mathcal{H}_2 approximation error.

REMARK 4.1. The similar phenomenon is observed in IRKA as well, see [7]. *To cope with this behavior, the outer loop that governs the VF (or IRKA) iterations must have memory and be equipped with a device capable of recognizing periodicity numerically (up to a tolerance).* Note that this is a more sophisticated control of the iterations, where periodicity is just one of many possible events that can be captured. For these types of iterations, the usual memoryless loop breaking (comparing only consecutive steps, or testing against a stopping criterion) is not enough. Instead, for instance, a loop control with memory can be used for early detection of upcoming numerical convergence and better steering of the iterations, see e.g. [22]. Clearly, if the poles enter a periodic behavior, the distance $\delta_k = \text{dist}(\lambda^{(k)}, \lambda^{(k-1)})$ will become periodic; and consequently it is enough to test the sequence (δ_k) for periodicity. If $\tau \geq 1$ is the estimated period, then we have τ candidate sets of poles $\lambda^{(k)}, \dots, \lambda^{(k+\tau-1)}$ for the approximation. If these poles are not satisfactory, the looping must be interrupted. Details are deferred to a subsequent work.

4.3. Avoiding ill-conditioning via regularization. The matrices \mathcal{B} in (2.3) and \mathcal{A} in (2.7) appearing in the SK and VF iterations, respectively, are composed of notoriously ill-conditioned Vandermonde and Cauchy matrices. For instance, the spectral condition number $\kappa_2(V) = \|V\|_2 \|V^{-1}\|_2$

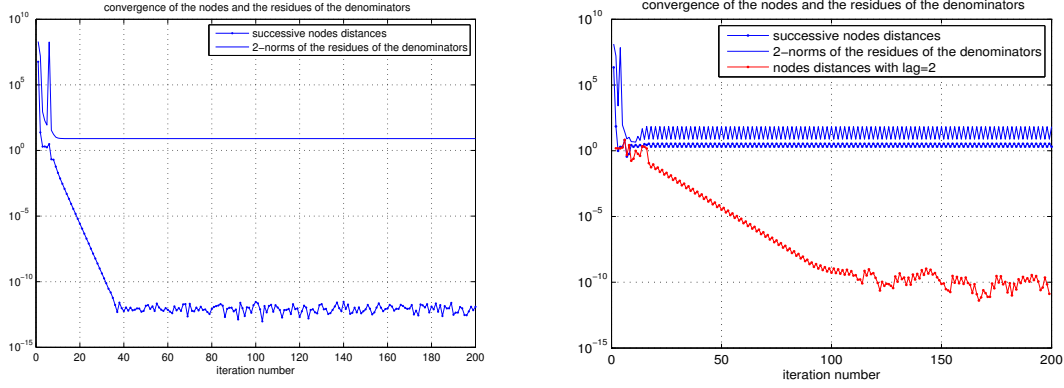


FIG. 4.1. *Left* ($r = 18$): The distance between consecutive sets of computed poles, $\|(\boldsymbol{\lambda}^{(k)} - \boldsymbol{\lambda}^{(k+1)})/\boldsymbol{\lambda}^{(k)}\|_\infty$ and $\|(\tilde{\varphi}_j^{(k)})_{j=1}^r\|_2$. *Right* ($r = 17$): The relative differences $\|(\boldsymbol{\lambda}^{(k)} - \boldsymbol{\lambda}^{(k+1)})/\boldsymbol{\lambda}^{(k)}\|_\infty$ (jumping between 2.02 and 3.41), $\|(\boldsymbol{\lambda}^{(k)} - \boldsymbol{\lambda}^{(k+2)})/\boldsymbol{\lambda}^{(k)}\|_\infty$, and the residues of the denominators $\|(\tilde{\varphi}_j^{(k)})_{j=1}^r\|_2$ (jumping between 7.51 and 68.66).

of an arbitrary real $n \times n$ Vandermonde matrix exceeds $2^{n-2}/\sqrt{n}$. For example, $\kappa_2(V) > 10^{28}$ for $n = 100$; see, e.g., [10, 52] for details. Cauchy matrices can also be similarly as badly conditioned. The Hilbert matrix is the most famous one; for the Hilbert matrix Hilb_{100} of order 100, $\kappa_2(\text{Hilb}_{100}) > 10^{150}$. In addition to being already ill-conditioned, the Cauchy matrices arising in VF appear with additional scalings, as $\Delta^{(k)}\mathcal{A}^{(k)} = \Delta^{(k)}(\mathcal{C}^{(k)} D_\xi \mathcal{C}^{(k)})$, where $\Delta^{(k)}$ is as defined in (2.3), $\mathcal{C}^{(k)}$ is a Cauchy matrix defined as $C_{ij} = \frac{1}{\xi_i - \lambda_j}$ for $i = 1, \dots, \ell$ and $j = 1, \dots, r$ and D_ξ is a diagonal matrix with $(D_\xi)_{kk} = -H(\xi_k)$ for $k = 1, \dots, \ell$. The diagonal matrices $\Delta^{(k)}$ and D_ξ can also be arbitrarily ill-conditioned. For instance, if $H(s)$ has a pole in the vicinity of ξ_j , then $|(D_\xi)_{jj}| = |H(\xi_j)|$ might be very large, especially much larger than $|H(\xi_i)|$ where $|\xi_i|$ is big and thus $|H(\xi_i)|$ is small since $H(s)$ is assumed to be strictly proper. Hence, the LS problem contains potentially extremely ill-conditioned coefficient matrices, and the normal equations approach, used in the early development of LS rational approximations [40, 49], is in general not feasible. One of the key improvements of VF [30] is indeed removing the scaling by $\Delta^{(k)}$ and using the unscaled matrix $\mathcal{A}^{(k+1)} = (\mathcal{C}^{(k+1)} D_\xi \mathcal{C}^{(k+1)})$ instead. However, this matrix still remains ill-conditioned.

Although VF and SK iteration perform effectively for smaller r values and for not-too-pathological distributions of nodes and poles, the high condition number of the underlying Cauchy and Vandermonde matrices has been recognized as a serious obstacle for robust computations with higher order approximants on wider frequency ranges. A discussion on how this ill-conditioning affects the quality of the approximation of the SK iterations, including illustrative examples, is given in [50], where the authors demonstrate that equilibrating the columns of the LS coefficient matrix in many cases dramatically improves the accuracy. Another similar preconditioning technique is frequency scaling proposed in [45]. In this section, we will propose a regularization-based approach to remedy ill-conditioning.

RegVF: Regularized Vector Fitting. Increasing the order of the approximant naturally increases the potential of better approximation, but unfortunately only in theory. To illustrate this point, we continue the numerical experiment of Example 3.3, use the *Beam* example and increase the order of the approximant from $r = 40$ to $r = 80$. Recall that in Example 3.3 with $r = 40$, VF leads to a relative \mathcal{H}_2 error of $4.84 \cdot 10^{-2}$. However, when we increase r to $r = 80$, the relative error \mathcal{H}_2 of VF increases to $1.31 \cdot 10^2$. Of course, this apparent numerical divergence is solely due to the ill-conditioned LS problems, and in this context even the dimension $r = 80$ can be considered large.

One possible cure is to regularize the solution. Towards this goal, we introduce the concept of *Regularized Vector Fitting*, RegVF. In RegVF, we augment the LS coefficient matrix and replace the original problem (2.12) with

$$\left\| \begin{pmatrix} \mathcal{A}^{(k+1)} \\ \eta_1 I_r & 0 \\ 0 & \eta_2 I_r \end{pmatrix} \tilde{x}^{(k+1)} - \begin{pmatrix} h \\ 0_{2r \times 1} \end{pmatrix} \right\|_2 \rightarrow \min, \quad k = 0, 1, 2, \dots, \quad (4.6)$$

where η_1 and η_2 are the appropriately chosen regularization parameters. In a similar manner, we also regularize the final LS solution for the residue identification step. For the Beam model with $r = 80$ and for the same nodes, this modification together with the choices of $\eta_1 = 10^{-16}$, $\eta_2 = \sqrt{\epsilon \rho s}$ reduces the relative error from $1.31 \cdot 10^2$ to $1.48 \cdot 10^{-3}$. Needless to say, finding optimal regularization parameters in practice is far from trivial, because the *backslash* LS solver and the `svd()` function in Matlab are not a match for the highly ill-conditioned Cauchy-type matrices. For the sake of brevity, we omit the details to be included in [21].

A note on row scaling. To remedy ill-conditioning, in addition to column scaling, Soysal and Semlyen [50] proposed other approaches such as frequency shifting and row scaling. We note that preconditioning by row scaling in the context of LS may not be allowed, because it overrides carefully determined row weighting of a quadrature formula (see §3.3, §3.5), or row scaling designed to cope with measurement noise, see §5. The ill-conditioning induced by row-weighting can be partially overcome if the QR factorization is computed with the full pivoting introduced by Powell and Reid [48] and analyzed by Cox and Higham [16]. Therefore, if row scaling is an issue, before using the *backslash* LS solver in VF, we propose the following equally good yet more efficient simplified variant of Powell-Reid pivoting, due to Åke Björck:

```
function x = LS_solve( A, b )
m = size(A,1) ; D = zeros(m,1) ; for i = 1 : m, D(i) = norm(A(i,:),inf) ; end
[~,P] = sort(D,'descend') ; A = A(P,:) ; b = b(P) ; x = A \ b ;
```

REMARK 4.2. It is well known that using orthonormal basis functions improves numerical stability of approximation methods. For rational approximation schemes such as VF, several authors have developed methods based on orthogonal rational functions, e.g., [1, 18]. Examples where the orthonormal vector fitting (OrthVF) can outperform VF are given in [2]. However, this is still an open debate as Gustavsen [28] points out that careful implementation of VF with suitably chosen initial poles matches the performances of OrthVF on the same examples used in [2].

5. Vector Fitting using noisy data. The starting point for the rational approximation framework we consider in this paper is a set of transfer function measurements/evaluations. Even though so far we have only considered noise-free data and even though a complete analysis of the underlying framework for VF in the presence of noise is not the main focus of this paper, in this section we provide a new formulation for VF for noisy data and illustrate that the pole reallocation feature of VF leads to a powerful mechanism for removing noise asymptotically as the iteration advances. We also propose a new numerical linear algebra framework for the noisy data case, and pose some challenging problems for future research.

5.1. A mixed total least squares framework. Suppose that in (1.1), instead of exact values $h_i = H(\xi_i)$, we have noisy measured data: $\tilde{h}_i = h_i + \delta h_i$. Assuming that measurement errors are uncorrelated, the proper formulation of the new LS problem is

$$\text{Find } H_r(s) = \frac{n(s)}{d(s)} \equiv \frac{\sum_{j=0}^{r-1} \alpha_j s^j}{1 + \sum_{j=1}^r \beta_j s^j} \text{ such that } \sum_{i=1}^{\ell} w_i^2 \left| \frac{n(\xi_i)}{d(\xi_i)} - \tilde{h}_i \right|^2 \rightarrow \min, \quad (5.1)$$

where the weight w_i is the reciprocal of the standard deviation for the i th measurement - information that can be considered as part of the measurement and is essential in guiding the approximation process. Neglecting the weights w_i corresponds to assuming the same variance across all measurements. Such an assumption is generally not realistic; particularly when the measurements, \tilde{h}_i , span a large range of values. This, in turn, will degrade the performance of VF, causing it hopelessly to try to fit the noise.

Statistical properties of the errors, δh_i , can generally be obtained through repeated measurements, e.g. with periodic excitation, and depending on the model (see e.g. [46, §IV.]), various formulations are obtained. In general, if we set

$$\boldsymbol{\epsilon} = [\epsilon_1, \epsilon_2, \dots, \epsilon_\ell]^T \quad \text{with} \quad \epsilon_i = \frac{n(\xi_i)}{d(\xi_i)} - \tilde{h}_i \quad \text{for } i = 1, \dots, \ell,$$

the problem (5.1) can be re-formulated as $\|W\boldsymbol{\epsilon}\|_2 \rightarrow \min$, where W is the inverse Cholesky factor of a positive definite variance-covariance matrix. We assume for simplicity that errors are uncorrelated so that W is diagonal, and that further the individual error variances can be estimated reliably. Here we focus on the numerical linear algebra aspects of the problem. For details on stochastic estimation of transfer functions, see, e.g., [37, 46, 47, 54].

The SK iteration, which forms the basis for VF, now takes the weighted form

$$\left\| W \text{diag}\left(\frac{1}{|d^{(k)}(\xi_i)|}\right) \begin{pmatrix} n^{(k+1)}(\xi_1) - d^{(k+1)}(\xi_1)\tilde{h}_1 \\ \vdots \\ n^{(k+1)}(\xi_\ell) - d^{(k+1)}(\xi_\ell)\tilde{h}_\ell \end{pmatrix} \right\|_2 \equiv \|WD^{(k)}(\tilde{\mathcal{A}}^{(k)}x^{(k+1)} - \tilde{h})\|_2 \longrightarrow \min, \quad k = 0, 1, \dots$$

In VF, due to pole relocation, the scaling factors $1/|d^{(k)}(\xi_i)|$ are dropped and $D^{(k)} \equiv I$. The LS objective is $\|W(\tilde{\mathcal{A}}^{(k+1)}\tilde{x}^{(k+1)} - \tilde{h})\|_2 \rightarrow \min$. To ease the growing notational burden, we drop the iteration index k and set $\mathcal{A} \equiv \tilde{\mathcal{A}}^{(k+1)}$ and $x \equiv \tilde{x}^{(k+1)}$. (If needed, we may assume that k is big enough, so that the VF iterations have reached numerical convergence.) Compare the original LS problem in (5.1) with this linearized version. Note that, in the process of linearization, noise that had appeared only in the right-hand side of (5.1) now enters the coefficient matrix, leading to the minimization problem: $\|W(\mathcal{A}x - (h + \delta h))\|_2 \rightarrow \min$, where $\mathcal{A} = \mathcal{A}_{\text{noise-free}} + \delta\mathcal{A}$.

Since the tacit assumption of LS approximation is that only the right-hand side is contaminated with noise, a Total Least-Squares (TLS) formulation [25] appears to be more appropriate to this setting than the more typical LS formulation. More precisely, allowing for noisy data in (2.6), (2.12), (2.2) will lead to mixed LS/TLS problems. Notice that from the definition (2.7) only the last r columns of the matrix \mathcal{A} can be contaminated by noise, since ξ_i and λ_j are considered exact. Thus, the perturbation $\delta\mathcal{A}$ due to noise is structured and closely related to the perturbation δh :

$$\delta\mathcal{A} = \begin{pmatrix} 0 & F \end{pmatrix}, \quad F = \left(\frac{\delta h_i}{\xi_i - \lambda_j} \right)_{i,j=1,1}^{\ell,r} \equiv (\delta h \mathbf{e}^T) \circ C, \quad C_{ij} = \frac{1}{\xi_i - \lambda_j}, \quad \mathbf{e}^T = (1 \ 1 \ \dots \ 1), \quad (5.2)$$

where \circ again denotes the Hadamard product. Note that F has rank-one displacement structure, i.e., it satisfies a Sylvester equation with a rank-one nonhomogeneity:

$$\Xi F - F \Lambda = \delta h \mathbf{e}^T, \quad \text{with } \Xi = \text{diag}(\xi_i)_{i=1}^\ell \quad \text{and} \quad \Lambda = \text{diag}(\lambda_j)_{j=1}^r.$$

We first consider how VF fits into this general TLS framework and show that pole-relocation that is intrinsic to VF has useful additional consequences in this setting. Recall that the minimization of $\|W(\mathcal{A}x - \tilde{h})\|_2$ can be equivalently formulated as

$$\|W\mathbf{r}\|_2 \rightarrow \min, \quad \text{subject to } \tilde{h} + \mathbf{r} \in \text{Range}(\mathcal{A}). \quad (5.3)$$

One may find the solution to (5.3) by seeking the minimal change, $\tilde{h} \mapsto \tilde{h} + \mathbf{r}$, (as measured in a W -weighted norm) such that $\mathcal{A}x = \tilde{h} + \mathbf{r}$. In the general TLS framework, a minimal change $(\Delta\mathcal{A} \ \hat{\mathbf{r}})$ is determined (as measured now by weighted matrix norm: $\|W(\Delta\mathcal{A} \ \mathbf{r})T\|_F$) such that $(\mathcal{A} + \Delta\mathcal{A})\hat{x} = \tilde{h} + \hat{\mathbf{r}}$. If the entries of $(\Delta\mathcal{A} \ \mathbf{r})$ are uncorrelated, then W and $T = \text{diag}(t_i)_{i=1}^{2r+1}$ are diagonal matrices. For a detailed and instructive discussion on scaling, see [34, §3.6.2].

When there is no structural requirement on the perturbation $\Delta\mathcal{A}$, the TLS solution is computed as follows [25]: Let $G \equiv W \begin{pmatrix} \mathcal{A} & \tilde{h} \end{pmatrix} T = U\Sigma V^*$ be the SVD, and assume for simplicity that the smallest singular value $\sigma_{2r+1} > 0$ is simple, with the corresponding singular vectors u_{2r+1} (left) and v_{2r+1} (right). Further assume that the last component of v_{2r+1} is nonzero; i.e. $v_{2r+1} = \begin{pmatrix} z \\ \eta \end{pmatrix}$ where $z \in \mathbb{C}^{2r}$, $\eta \in \mathbb{C}$ with $\eta \neq 0$. Then, the minimal perturbation $(\Delta\mathcal{A} \ \hat{\mathbf{r}})$ and the corresponding solution \hat{x} are given explicitly as

$$(\Delta\mathcal{A} \ \hat{\mathbf{r}}) = -\sigma_{2r+1} W^{-1} u_{2r+1} v_{2r+1}^* T^{-1}, \quad \hat{x} = \frac{-1}{\eta t_{n+1}} \text{diag}(t_i)_{i=1}^{2r} z. \quad (5.4)$$

For more details on the solution procedure see [34, Algorithm 3.1].

Considering the special structure (5.2) of the perturbation in our setting, we formulate the following structured mixed LS/TLS problem: With W as before and $T = \text{diag}(t_i)_{i=1}^{r+1}$, solve

$$\|W(E \ \mathbf{r})T\|_F \longrightarrow \min, \text{ subject to } h + \mathbf{r} \in \text{Range}(\mathcal{A} + (0 \ E)) \text{ and } \Xi E - E\Lambda = \mathbf{r}e^T. \quad (5.5)$$

Since $W(E \ \mathbf{r})T = (T \otimes W) \begin{pmatrix} \text{vec}(E) \\ \mathbf{r} \end{pmatrix}$, the objective function in (5.5) can be re-written as

$$\|W(E \ \mathbf{r})T\|_F^2 = \left\| \begin{pmatrix} t_1 W & & \\ & \ddots & \\ & & t_{r+1} W \end{pmatrix} \begin{pmatrix} \mathbf{r} \circ C(:,1) \\ \vdots \\ \mathbf{r} \circ C(:,r) \end{pmatrix} \right\|_F^2 = t_{r+1}^2 \|W\mathbf{r}\|_2^2 + \sum_{j=1}^r t_j^2 \|W(\mathbf{r} \circ C(:,j))\|_2^2. \quad (5.6)$$

Depending on T and the distribution of the ξ_i 's and the λ_j 's, minimizing the above expression is related to minimizing $\|W\mathbf{r}\|_2$. For example consider the case of the structured perturbations $E = \mathbf{r}e^T$ (not of the type we have here¹, but instructive to consider) and $T = I$ (reasonable in this situation). In this case, the minimization problem $\|W(E \ \mathbf{r})T\|_F \longrightarrow \min$ is indeed equivalent to $\|W\mathbf{r}\|_2 \longrightarrow \min$. But in general, developing a theory for solvability and a robust numerical algorithm for solving (5.5) is a challenging problem. If we assume to have found the minimizing E and $\hat{\mathbf{r}}$, the solution \hat{x} is, then, defined by $(\mathcal{A} + (0 \ E))\hat{x} = h + \hat{\mathbf{r}}$. Otherwise (e.g., if (5.5) has no solution), ignore the rank-one displacement structure, and use the solution of the mixed LS/TLS problem, computed using [34, Algorithm 3.2], or the solution (5.4) of the TLS problem. With these two cases, we obtain two new variants of VF, denoted by LS/TLS-VF and TLS-VF, respectively. As stated above, the special structure of the coefficients matrices in LS/TLS-VF makes it a challenging problem. Assuming the existence of a solution to these structured problems, numerically sound implementations to obtain the solution will depend on developing accurate numerical linear algebra tools, e.g., accurate SVD computations, for Cauchy-type matrices that arise in VF.

5.2. VF as an asymptotic LS/TLS procedure. We compare the LS solution in step k of VF to the solution of $(\mathcal{A} + (0 \ E))\hat{x} = h + \hat{\mathbf{r}}$ in step k of LS/TLS-VF. Recall that the LS problem minimizes $\|W\mathbf{r}\|_2$, and the solution x satisfies $\mathcal{A}x = h + \mathbf{r}$. If we partition x as $x = \begin{pmatrix} \phi \\ \varphi \end{pmatrix}$, and add the LS/TLS error in \mathcal{A} , we obtain

$$(\mathcal{A} + (0 \ E))x = h + \mathbf{r} + E\varphi,$$

¹Take very low frequencies and all λ_j 's around -1 , or consider frequency scaling to approximate the desired structure.

where in the case of (5.5), $E\varphi = ((\hat{\mathbf{r}}e^T) \circ C)\varphi = \hat{\mathbf{r}} \circ (C\varphi)$. Since the φ -part of x in VF converges to zero, it holds that $\|E\varphi\|_2 \leq \|\hat{\mathbf{r}}\|_2 \|C\varphi\|_2$ is small relative to $\|\hat{\mathbf{r}}\|_2$. Recall that minimizing $\|W\mathbf{r}\|_2$ and $\|W(E - \mathbf{r})\|_2$ from (5.6) are related. This reveals another silent, yet powerful, feature of VF that makes it much more than a reformulation of SK iteration. Asymptotically, thanks to the persistent change of representation through pole relocation, VF is (approximately) performing structured mixed LS/TLS minimization.

5.3. A diagonally-restricted LS/TLS formulation. In the previous section, we discussed the TLS approach to VF in the presence of noise and by comparing with a generic LS/TLS procedure we showed that the original formulation of VF will approximately solve the LS/TLS problem. In this section, we will introduce a new framework, that we believe is the correct formulation to perform VF in the presence of noise.

It follows from (5.3) and the definition of C in (5.2) that the objective function $\|W\mathbf{r}\|_2$ is minimized with respect to the condition

$$((C \text{ diag}(\tilde{h})C) + (0 \text{ diag}(\mathbf{r})C))x = \tilde{h} + \mathbf{r}, \text{ i.e. } ((C \text{ diag}(\tilde{h})C \text{ } \tilde{h}) + \text{diag}(\mathbf{r}) (0 \text{ } C \text{ } \mathbf{e})) \begin{pmatrix} x \\ -1 \end{pmatrix} = 0.$$

Define $Z = (C \text{ diag}(\tilde{h})C \text{ } \tilde{h})$, $S = (0 \text{ } C \text{ } \mathbf{e})$. Then we propose a diagonally-restricted LS/TLS formulation in step k of VF, stated as follows: Find $x = \begin{pmatrix} \phi \\ \varphi \end{pmatrix}$ as the solution (if it exists) of the constrained minimization problem

$$\min_{\mathbf{r}} \{\|W\mathbf{r}\|_2 : (Z + \text{diag}(\mathbf{r})S) \begin{pmatrix} x \\ -1 \end{pmatrix} = 0\}. \quad (5.7)$$

Set $\hat{Z} = WZ$ and note that $\hat{R} \equiv W\text{diag}(\mathbf{r})$ is the minimal perturbation $\hat{Z} \rightsquigarrow \hat{Z} + \hat{R}S$ that makes \hat{Z} singular. Apart from the special structure of \hat{R} , this is related to the notion of restricted singular values [56] of the matrix triplet (\hat{Z}, I, S) : $\sigma_k(Z, I, S) = \min\{\|\Theta\|_2 : \text{rank}(Z + I\Theta S) \leq k - 1\}$. This connection, explained in [53], together with methods presented in [9] form the starting point for attacking the problem of solving (5.7) numerically. These issues will be explored in future work.

6. Conclusions and Future Directions. VF has been widely and successfully used. Notwithstanding substantial advances and many successful applications of the method, analytical justification of its success from numerical linear algebra and rational approximation perspectives has been missing. This work is a step toward filling that gap. Noting first that a small VF fitting error does not necessarily correspond to small approximation error, we related VF to discrete \mathcal{H}_2 minimization and proposed a quadrature-based version, called QuadVF, which improves performance dramatically. We extended VF to include a derivative penalty in the LS minimization by performing a quadrature-based discretization of a continuous Sobolev norm, leading to a method we called SobVF. We also analyzed several practical and numerical issues arising in VF using a rigorous theoretical framework. For example, we analytically justified the mechanism behind the mirroring of unstable poles during VF. We investigated the numerical convergence of VF and illustrated different scenarios for divergence that could arise. One of the major numerical issues that can arise in VF is the appearance of highly ill-conditioned coefficient matrices; we offered a remedy via regularization. Even though most of our analyses assume exact data, we briefly considered VF in the case of noisy data and showed the utility of a mixed LS/TLS framework.

Aside from the newly developed, effective methods that are described here, our work also leads to a variety of challenging theoretical and practical issues that will be explored in subsequent work. These include: effective regularization techniques, refined computational strategies for the diagonally-restricted LS/TLS formulation of VF introduced in (5.7), extensions to the multiple-input/multiple-output case via tangential interpolation (reflecting the structure of the underlying \mathcal{H}_2 setting), and adaptive determination of appropriate reduced dimension (say, informed by the Loewner framework developed in [4, 38, 41, 44]).

REFERENCES

- [1] H. AKÇAY AND B. NINNESS, *Orthonormal basis functions for modelling continuous-time systems*, Signal Processing, 77 (1999), pp. 261–274.
- [2] G. ANTONINI, D. DESCHRIJVER, AND T. DHAENE, *A comparative study of vector fitting and orthonormal vector fitting techniques for EMC applications*, in Proc. Int. Symp. Electromagnetic Compatibility, IEEE, 2006, pp. 6–11.
- [3] A. ANTOULAS, C. BEATTIE, AND S. GUGERCIN, *Interpolatory model reduction of large-scale dynamical systems*, in Efficient Modeling and Control of Large-Scale Systems, J. Mohammadpour and K. Grigoriadis, eds., Springer-Verlag, 2010, pp. 2–58.
- [4] A. ANTOULAS, A. IONITA, AND S. LEFTERIU, *On two-variable rational interpolation*, Linear Algebra and Its Applications, 436 (2012), pp. 2889–2915.
- [5] A. ANTOULAS, D. SORENSSEN, AND S. GUGERCIN, *A survey of model reduction methods for large scale systems*, Contemporary Mathematics, AMS Publications, 280 (2001), pp. 193–219.
- [6] C. BEATTIE, Z. DRMAČ, AND S. GUGERCIN, *A note on shifted Hessenberg systems and frequency response computation*, ACM Trans. Math. Softw., 38 (2012), pp. 12:1–12:16.
- [7] C. BEATTIE, Z. DRMAČ, AND S. GUGERCIN, *A reproducing kernel framework for optimal \mathcal{H}_2 model order reduction*, tech. rep., University of Zagreb and Virginia Tech at Blacksburg, 2013.
- [8] C. BEATTIE AND S. GUGERCIN, *Realization-independent \mathcal{H}_2 approximation*, in Proceedings of the 51st IEEE Conference on Decision & Control, IEEE, 2012, pp. 4953–4958.
- [9] A. BECK, *The matrix-restricted total least-squares problem*, Signal Process., 87 (2007), pp. 2303–2312.
- [10] B. BECKERMANN, *The condition number of real Vandermonde, Krylov and positive definite Hankel matrices*, Numer. Math., 85 (2000), pp. 553–577.
- [11] R. BHATIA, *Matrix Analysis*, Springer, 1997. Graduate Texts in Mathematics, 169.
- [12] J. P. BOYD, *The optimization of convergence for chebyshev polynomial methods in an unbounded domain*, Journal of computational physics, 45 (1982), pp. 43–79.
- [13] J. P. BOYD, *Exponentially convergent Fourier-Chebyshev quadrature schemes on bounded and infinite intervals*, Journal on Scientific Computing, 2 (1987), pp. 99–109.
- [14] Y. CHAHLAOUI AND P. V. DOOREN, *A collection of benchmark examples for model reduction of linear time invariant dynamical systems*, tech. rep., SLICOT Working Note 2002-2, 202.
- [15] H. CHEN, J. ZHENG, AND J. FANG, *Multipoint moment matching based model generation for complex systems*, in Electrical Performance of Electronic Packaging (Princeton, NJ), IEEE, 2003, pp. 299–302.
- [16] A. J. COX AND N. J. HIGHAM, *Stability of Householder QR factorization for weighted least squares problems*, in Numerical Analysis 1997, Proceedings of the 17th Dundee Biennial Conference, D. F. Griffiths, D. J. Higham, and G. A. Watson, eds., vol. 380 of Pitman Research Notes in Mathematics, A W Longman, 1998, pp. 57–73.
- [17] D. DESCHRIJVER AND B. GUSTAVSEN, *Advancements in iterative methods for rational approximation in the frequency domain*, IEEE Transactions on Power Delivery, 22 (2007), pp. 1633–1642.
- [18] D. DESCHRIJVER, B. HAEGEMAN, AND T. DHAENE, *Orthonormal vector fitting: a robust macromodeling tool for rational approximation of frequency domain responses*, IEEE Transactions on Advanced Packaging, 30 (2007), pp. 216–225.
- [19] D. DESCHRIJVER, L. KNOCKAERT, AND T. DHAENE, *Improving robustness of vector fitting to outliers in data*, Electronics Letters, 46 (2010), pp. 1–2.
- [20] D. DESCHRIJVER, M. MROZOWSKI, T. DHAENE, AND D. D. ZUTTER, *Macromodeling of multiport systems using a fast implementation of the vector fitting method*, IEEE Microwave and Wireless Components Letters, 18 (2008), pp. 383–385.
- [21] Z. DRMAČ, *Accurate SVD of Cauchy-type matrices and applications*, tech. rep., University of Zagreb, 2014.
- [22] Z. DRMAČ AND K. VESELIĆ, *New fast and accurate Jacobi SVD algorithm: II.*, SIAM J. Matrix Anal. Appl., 29 (2008), pp. 1343–1362.
- [23] P. FELDMANN AND R. W. FREUND, *Efficient linear circuit analysis by Pade approximation via the Lanczos process*, Trans. Comp.-Aided Des. Integ. Cir. Sys., 14 (1995), pp. 639–649.
- [24] W. GAUTSCHI, *Orthogonal Polynomials, Computation and Approximation*, Oxford University Press, 2004. Numerical Mathematics and Scientific Computation.
- [25] G. H. GOLUB AND C. F. V. LOAN, *An analysis of the total least squares problem*, SIAM J. Numer. Anal., 17 (1980), pp. 214–224.
- [26] S. GUGERCIN, A. ANTOULAS, AND N. BEDROSSIAN, *Approximation of the international space station 1r and 12a models*, in Decision and Control, 2001. Proceedings of the 40th IEEE Conference on, vol. 2, IEEE, 2001, pp. 1515–1516.
- [27] S. GUGERCIN, A. C. ANTOULAS, AND C. BEATTIE, *\mathcal{H}_2 model reduction for large-scale linear dynamical systems*, SIAM J. Matrix Anal. Appl., 30 (2008), pp. 609–638.
- [28] B. GUSTAVSEN, *Comments on "a comparative study of vector fitting and orthonormal vector fitting techniques for emc applications"*, in Proceedings of the 18th Int. Zurich Symposium on Electromagnetic Compatibility,

- Munich 2007, IEEE, 2006, pp. 131–134.
- [29] ———, *Improving the pole relocating properties of vector fitting*, IEEE Transactions on Power Delivery, 21 (2006), pp. 1587–1592.
 - [30] B. GUSTAVSEN AND A. SEMLYEN, *Rational approximation of frequency domain responses by vector fitting*, IEEE Transactions on Power Delivery, 14 (1999), pp. 1052–1061.
 - [31] W. HENDRICKX, D. DESCHRIJVER, AND T. DHAENE, *Some remarks on the Vector Fitting iteration*, in in Mathematics in Industry, Springer-Verlag, 2006, pp. 134–138.
 - [32] W. HENDRICKX AND T. DHAENE, *A discussion of "Rational approximation of frequency domain responses by vector fitting"*, IEEE Transactions on Power Systems, 21 (2006), pp. 441–443.
 - [33] A. HOCHMAN, Y. LEVIATAN, AND J. WHITE, *On the use of rational-function fitting methods for the solution of 2D Laplace boundary-value problems*, tech. rep., arXiv:1112.1643v2, 2012.
 - [34] S. V. HUFFEL AND J. VANDEWALLE, *The Total Least Squares Problem: Computational Aspects and Analysis*, Frontiers in Applied Mathematics 9, SIAM, 1991.
 - [35] R. E. KALMAN, *Design of a self-optimizing control system*, Trans. ASME, 80 (1958), pp. 468–478.
 - [36] K. J. KIM, R. COOLS, AND L. G. IXARU, *Quadrature rules using first derivatives for oscillatory integrands*, J. Comput. Appl. Math., 140 (2002), pp. 479–497.
 - [37] I. KOLLÁR, *On frequency-domain identification of linear systems*, IEEE Transactions on Instrumentation and Measurement, 42 (1993), pp. 2–6.
 - [38] S. LEFTERIU AND A. ANTOULAS, *A new approach to modeling multiport systems from frequency-domain data*, Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on, 29 (2010), pp. 14–27.
 - [39] S. LEFTERIU AND A. C. ANTOULAS, *Convergence of the vector fitting algorithm*, IEEE Transactions on Microwave Theory and Techniques, 61 (2013), pp. 1435–1443.
 - [40] E. C. LEVY, *Complex curve fitting*, IRE Transactions on Automatic Control, AC-4 (1959), pp. 37–44.
 - [41] A. J. MAYO AND A. C. ANTOULAS, *A framework for the solution of the generalized realization problem*, Linear Algebra and its Applications, 425 (2007), pp. 634–662.
 - [42] V. MEHRMANN AND H. XU, *An analysis of the pole placement problem. I. the single-input case*, Electronic Transactions on Numerical Analysis, 4 (1996), pp. 89–105.
 - [43] L. MEIER AND D. G. LUENBERGER, *Approximation of linear constant systems*, IEEE Trans. Automat. Control, 12 (1967), pp. 585–588.
 - [44] A. ONITA AND A. ANTOULAS, *Data-driven parametrized model reduction in the Loewner framework*. submitted, 2013.
 - [45] R. PINTELON AND I. KOLLR, *On the frequency scaling in continuous-time modeling.*, IEEE T. Instrumentation and Measurement, 54 (2005), pp. 318–321.
 - [46] R. PINTELON, Y. ROLAIN, J. SCHOUKENS, AND H. V. HAMME, *Parametric identification of transfer functions in the frequency domain - a survey*, IEEE Transactions on Automatic Control, 39 (1994), pp. 2245–2260.
 - [47] R. PINTELON, J. SCHOUKENS, AND Y. ROLAIN, *Uncertainty of transfer function modeling using prior estimated noise models*, in 13 IFAC Symposium on System Identification, Elsevier, 2003, pp. 1874–1879.
 - [48] M. J. D. POWELL AND J. K. REID, *On applying Householder transformations to linear least squares problems*, in Information Processing 68, Proc. International Federation of Information Processing Congress, Edinburgh, 1968, North Holland, Amsterdam, 1969, pp. 122–126.
 - [49] C. SANATHANAN AND J. KOERNER, *Transfer function synthesis as a ratio of two complex polynomials*, IEEE Trans. Autom. Control, 8 (1963), pp. 56–58.
 - [50] A. O. SOYSAL AND A. SEMLYEN, *Practical transfer function estimation and its applications to wide frequency range representation of transformers*, IEEE Transactions on Power Delivery, 8 (1993), pp. 1627–1637.
 - [51] J. SPANOS, M. MILMAN, AND D. MINGORI, *A new algorithm for L^2 optimal model reduction*, Automatica (Journal of IFAC), 28 (1992), pp. 897–909.
 - [52] E. V. TYRTYSHNIKOV, *How bad are Hankel matrices?*, Numer. Math., 67 (1994), pp. 261–269.
 - [53] S. VAN HUFFEL AND H. ZHA, *The restricted total least squares problem: Formulation, algorithm, and properties*, SIAM J. Matrix Anal. Appl., 12 (1991), pp. 292–309.
 - [54] D. K. D. VRIES AND P. M. J. V. DEN HOF, *Quantification of uncertainty in transfer function estimation: a mixed-probabilistic-worst-case approach*, Automatica, 31 (1995), pp. 543–557.
 - [55] D. WILSON, *Optimum solution of model-reduction problem*, Proc. IEE, 117 (1970), pp. 1161–1165.
 - [56] H. ZHA, *The restricted singular value decomposition of matrix triplets*, SIAM J. Matrix Anal. Appl., 12 (1991), pp. 172–194.